languages

Article

Non-Word Repetition and Vocabulary in Arabic-Swedish-Speaking 4–7-Year-Olds with and without Developmental Language Disorder

Linnéa Öberg and Ute Bohnacker *

Department of Linguistics & Philology, Uppsala University, Box 635, SE-75126 Uppsala, Sweden; linnea.oberg@lingfil.uu.se

* Correspondence: ute.bohnacker@lingfil.uu.se

Abstract: The Arabic-speaking community in Sweden is large and diverse, yet linguistic reference data are lacking for Arabic-Swedish-speaking children. This study presents reference data from 99 TD children aged 4;0–7;11 on receptive and expressive vocabulary in the minority and the majority language, as well as for three types of non-word repetition (NWR) tasks. Vocabulary scores were investigated in relation to age, language exposure, and socio-economic status (SES). NWR performance was explored in relation to age, type of task, item properties, language exposure, and vocabulary. Eleven children with DLD were compared to the TD group. Age and language exposure were important predictors of vocabulary scores in both languages, but SES did not affect vocabulary scores in any language. Age and vocabulary size had a positive effect on NWR accuracy, whilst increasing item length and presence of clusters had an adverse effect. There was substantial overlap between the TD and DLD children for both vocabulary and NWR performance. Diagnostic accuracy was at best suggestive for NWR; no task or type of item was better at separating the two groups. Reports from parents and teachers on developmental history, language exposure, and functional language skills emerged as important factors for correctly identifying DLD in bilinguals.

Keywords: non-word repetition; NWR; vocabulary; bilingualism; preschool children; Arabic; Swedish; developmental language disorder; DLD

check for updates

Citation: Öberg, Linnéa, and Ute Bohnacker. 2022. Non-Word Repetition and Vocabulary in Arabic-Swedish-Speaking 4–7-Year-Olds with and without Developmental Language Disorder. *Languages* 7: 204. https://doi.org/10.3390/languages7030204

Academic Editor: Cornelia Hamann

Received: 31 March 2022
Accepted: 15 July 2022
Published: 2 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

This study investigates non-word repetition and vocabulary in a large group of bilingual Arabic-Swedish-speaking children with typical language development, compared to a smaller group of children with a DLD diagnosis. A large proportion of children in Sweden today grow up in a bilingual setting. According to official statistics (2022), 29% of all school children age 7–15 are entitled to mother tongue instruction, which means that they speak a home language other than (or in addition to) Swedish. During the past decades, the number of Arabic speakers has increased substantially, and Arabic is now considered to be the language with the second-highest number of native speakers in the country, after Swedish (National Agency for Education 2022; Parkvall 2016). Despite the fact that as many as a quarter of all Swedish children are bilingual, there is a lack of large-scale studies that investigate these children's language skills in both languages.

Developmental Language Disorder (DLD) is a common condition in children that negatively affects their oral communication, literacy and educational progress (Norbury et al. 2016).[1] DLD typically emerges in early childhood and manifests as a pronounced deficit in the development of language skills, which cannot be attributed to hearing impairment, intellectual disability, medical syndromes or neurological disorders (Bishop 1997, pp. 21–23; Leonard 2014, p. 3). Uncertainty about what should be considered 'normal' language development in bilingual children can lead to both over- and underdiagnosis of developmental disorders of language and literacy (Dollaghan and Horner 2011; Grimm and Schulz 2014).

More than two decades ago, an epidemiological study found that bilingual children in Sweden were referred to a Speech and Language Pathologist (SLP) for assessment at a later age than monolinguals, and they were also more likely to be considered to have severe DLD (Salameh et al. 2002). More recently, a very high proportion (82%) of Swedish child healthcare nurses have been found to believe that bilingualism causes language delay, and these nurses were more inclined to simplify screening and delay referrals for bilinguals (Nayeb et al. 2015). In a survey investigating the prevalence of *severe* DLD in five regions of the national healthcare service in Sweden, bilinguals were heavily overrepresented (51%) and bilingualism was reported to be a confounding factor, making it difficult for SLPs to make clinical judgments about the presence and severity of DLD (SOU 2016). This confusion can largely be attributed to insufficient assessment materials, a lack of reference data and patchy knowledge about developmental trajectories in bilinguals (Letts 2013). Furthermore, overlap in many of the linguistic features that are associated with DLD on the one hand and common patterns in typical L2 acquisition on the other adds to this confusion (Boerma et al. 2017a; Paradis and Crago 2000).

Although recommendations abound that bilinguals with suspected DLD should be assessed in both languages (ASHA 2004; World Health Organization 1992), and although it is frequently argued that DLD must manifest in *both* languages in bilinguals for a child to qualify for a diagnosis (Kohnert 2010; Salameh et al. 2002; Thordardottir 2015, p. 349), evidence-based recommendations about how to *interpret* language test scores for bilinguals are rare (Peña et al. 2016).

Typically, language test scores are converted into a standardised score in order to be able to compare the performance of an individual child against a reference/norm group. If performance is below a certain cut-off, this is interpreted as a language deficit and may lead to a diagnosis of DLD. Different countries have different clinical practices for assessing and identifying DLD, for instance regarding which cut-offs are utilised. As Thordardottir (2015) reports, clinical guidelines for diagnosing DLD in monolinguals in different European countries range between -2 (identifying the lowest-scoring 2.3%) to -1 (identifying the lowest-scoring 15.9%) z-scores below the mean on standardised language tests. Two large-scale epidemiological studies investigating the prevalence of DLD in monolingual children have proposed cut-offs of -1.25 (Tomblin et al. 1997) and -1.5 (Norbury et al. 2016) for composite language scores in a language domain or modality as a yardstick for diagnosis.

1.1. Vocabulary

Vocabulary is a cornerstone of general language skills and important for later academic achievement. Vocabulary is a linguistic domain that is maximally influenced by quantitative as well as qualitative aspects of language input. More exposure (child-directed speech) is associated with larger vocabularies and steeper vocabulary growth curves in children (Hart and Risley 1995; Rowe 2012). Qualitative aspects of the input, such as variation in syntax and rich vocabulary in child-directed speech and communication styles that are conducive to verbal interaction between adult and child, show positive effects on children's vocabulary growth (Cartmill et al. 2013; Rowe 2012). For bilingual children, language input is more variable, both concerning the amount of exposure to each language and the contexts and sources of such input (Paradis and Grüter 2014). As children grow older, their receptive and expressive vocabulary grows too (Haman et al. 2017), but in bilingual children this may not happen to the same extent in both languages. While many studies find that bilinguals increase their vocabulary scores in the majority language over time, vocabulary in the minority language may not increase to the same extent, or may even stagnate (Cobo-Lewis et al. 2002a, 2002b; Gagarina et al. 2014; Ganuza and Hedman 2019; Gathercole and Thomas 2009; Lindgren and Bohnacker 2020; Öztekin 2019). Frequently reported in the literature is also the influence of socio-economic status (SES) on vocabulary scores. At group level, bilingual children from families with high SES have been found to perform better on vocabulary tests in the majority language than children from families with low SES (Buac et al. 2014; Calvo and Bialystok 2014; Cobo-Lewis et al. 2002a; Gathercole et al. 2016; Leseman 2000; Prevoo et al. 2014). The effect of SES on the minority language is less consistent. While Cobo-Lewis et al. (2002b) found that children from low SES families perform better than children from high SES families on certain vocabulary tasks, other studies have not found an effect of SES in the minority language (Buac et al. 2014; Leseman 2000; Prevoo et al. 2014).

Children with DLD often have deficits in the lexical domain, with a slower rate of vocabulary growth (Rice and Hoffman 2015; Smolander et al. 2021) and smaller vocabularies than their typically developing peers. Such deficits in the lexical domain have been described for both monolingual and bilingual children with DLD (Boerma et al. 2017b; Khoury Aouad Saliby et al. 2017b; Spaulding et al. 2013; Thordardottir and Brandeker 2013).

At the same time, bilingual children with typical language development may have smaller vocabularies compared to monolinguals in one of their languages or both, depending on the relative amount of exposure to each language (Thordardottir 2011). Relative amount of exposure has been identified in several studies as a key predictor of majority and minority language vocabulary size (Prevoo et al. 2014; Unsworth 2016). Furthermore, the timing of the onset of bilingualism has also been investigated in relation to vocabulary development. Studies in this area generally find that while a binary categorisation of age of onset as early vs. late in itself is not a significant predictor of vocabulary scores later in life (Thordardottir 2011; Unsworth 2016), length of exposure (treated as a continuous variable) affects vocabulary size, with longer exposure times being associated with higher vocabulary scores. The association between length of exposure and vocabulary scores is modulated by the relative amount of exposure from age of onset to age at assessment, often referred to as cumulative exposure (Smolander et al. 2021; Thordardottir 2019).

In sum, vocabulary is affected by both bilingualism (due to variability in language exposure), and DLD. If a bilingual child scores low on vocabulary tasks in one language or both, it may be difficult to determine whether this is due to little exposure or due to DLD. Since vocabulary is probably the linguistic domain that is the most input-dependent, differences in exposure are likely to be reflected in unevenly sized vocabularies in each language. Moreover, it is frequently reported that typically developing bilinguals who are only assessed in the majority language perform significantly below the monolingual norm on standardised language tests targeting vocabulary (Boerma et al. 2017b; Peña et al. 2016) as well as general language skills (Andersson et al. 2019). By contrast, non-word repetition (NWR), which is discussed in the next section, is a task that has been said to be suitable for children of diverse cultural and linguistic backgrounds, as it is less biased than standardised language tests (Dollaghan and Campbell 1998; Thordardottir and Brandeker 2013).

1.2. NWR as a Diagnostic Tool for Identifying DLD

Non-word repetition is a task that entails imitating a sequence of phonological nonsense forms (non-words). Poor NWR performance has been known for over three decades to be a clinical marker of DLD in monolinguals in many different languages (Chiat 2015). For bilinguals as well, NWR has been described as a promising diagnostic tool. A number of studies have reported that poor NWR performance in bilingual children is an indicator of DLD (Boerma et al. 2015; de Almeida et al. 2017; Hamann and Abed Ibrahim 2017). Other work however has raised doubts as to whether NWR can reliably be used clinically for identifying DLD in bilingual children (Gutiérrez-Clellen and Simon-Cereijido 2010; Kohnert et al. 2006; Ortiz 2021).

Compared to other language measures, NWR is relatively little affected by language exposure, as it does not depend directly on language knowledge but rather on the processing of new language information (Archibald 2008). However, NWR performance is affected by a number of factors related to the characteristics of the non-words as well as by participant-related factors (for an overview, see Chiat 2015). For instance, item length (operationalised as number of syllables) and phonological complexity have been reported to affect repetition accuracy, where items generally become more difficult to repeat as length and complexity increases (Boerma et al. 2015; dos Santos and Ferré 2018; Ellis Weismer et al. 2000; Jones et al. 2010; Radeborg et al. 2006; Thordardottir and Brandeker 2013). Although not as well-studied, phonotactic probability, word-likeness, and prosodic features are also reported to affect NWR performance. NWR items with lower phonotactic probability, items carrying prosodic features with lower saliency, and items with a lower degree of word-likeness are typically more difficult to repeat (Chiat and Roy 2007; Gathercole 1995; Jones et al. 2010; Sahlén et al. 1999). Participant-related factors that influence NWR performance are, for instance, chronological age and lexical knowledge. NWR performance typically increases with age. The association between NWR performance and vocabulary size has been well known for several decades (Coady and Evans 2008). The relationship is likely to be bidirectional, meaning that better NWR capabilities facilitate vocabulary learning and that having a larger vocabulary facilitates NWR performance (Gathercole 2006). Several studies with bilingual participants also report an association between language exposure and performance on NWR tasks, especially when the items have language-specific features (Gibson et al. 2015; Kohnert et al. 2006; Sorenson Duncan and Paradis 2016; Thordardottir and Brandeker 2013). In light of this, some researchers have suggested that language-specific NWR tasks are unsuitable to use with bilinguals, and that NWR tasks that are constructed to be compatible with the phonological structure of many different languages may be better suited to identify DLD in bilinguals (Boerma et al. 2015).

Keeping in mind the influence of language exposure and vocabulary on repetition accuracy in language-specific tasks, a framework for constructing NWR tasks with different properties was developed within the COST Action IS0804 research network (Chiat 2015). In this framework, two main types of tasks are contrasted against the kind of language-specific tasks (LS) that have traditionally been used in NWR assessment (Gathercole et al. 1994; Radeborg et al. 2006). The first type is the so-called 'crosslinguistic' task (CL) with 2–5-syllable items and simple syllable structure (i.e., consonant-vowel syllables with no clusters or coda) that are constructed to be compatible with the phonological structure of many different languages (Chiat 2015; Boerma et al. 2015).[2] The second type is the so-called 'quasi-universal' task (QU), which has items with 1–3 syllables of varying syllabic complexity (clusters and codas), and probes phonological complexity (dos Santos and Ferré 2018).[3] In the present study, all three types of NWR tasks (LS, CL and QU) are used.

Information about early language development, risk factors of developmental disorders of language and literacy, as well as parental reports about functional language abilities may be useful in addition to standardised language tests when diagnosing DLD, particularly in bilinguals (Thordardottir 2015; Tuller 2015). A late emergence of the first word or the first multi-word utterance is associated with a greater risk of developing persistent language disorder later in life (Paradis et al. 2010; Trauner et al. 2000). At the same time, bilinguals with typical language development are expected to reach these early milestones at the same time as monolinguals, although they might not appear at the same time in both languages (Hoff et al. 2014). A family history of speech, language or literacy difficulties has been identified as a risk factor for DLD in both monolinguals and bilinguals (Kalnak et al. 2012; Restrepo 1998). In addition to parental reports, useful information about the child's language and communication can also be obtained from teachers and preschool staff. Teachers see the child every day, know about their learning outcomes, and observe them in interaction with peers and adults. Thus, teacher evaluations provide ecologically valid reports of children's functional language skills. Teacher descriptions of children's language abilities have been found to correlate with results on standardised language tests, and may also reveal language difficulties that are not always straightforwardly captured by standardised language tests (Botting et al. 1997; Purse and Gardner 2013).

1.3. The Present Study

Although there are many Arabic-Swedish-speaking children in Sweden, little is known about their language skills. Published studies are generally limited to certain aspects of morphosyntax and word-association in small groups of children (e.g., Holmström 2015, Salameh et al. 2004 and Håkansson et al. 2003). There is still a lack of large-scale studies that investigate both the majority and minority language and that also take into account age and environmental factors such as language exposure and SES. Furthermore, there is hardly any research on the NWR performance of bilingual children in Sweden. The present study aims to address this knowledge gap, by presenting reference data for vocabulary in both the minority and majority language and for three types of NWR tasks for a large sample (99 TD children) of Arabic-Swedish-speaking bilinguals aged 4–7. The relative effect of age, language exposure, and SES on vocabulary comprehension and production is investigated for both the majority and the minority language. Additionally, NWR performance is investigated in relation to age, language exposure, vocabulary and properties of the non-word items. Finally, this study explores whether bilingual children with a diagnosis of DLD can be distinguished from children with typical language development, based on their performance on vocabulary and NWR tasks. The following research questions are posed:

RQ1: How does vocabulary comprehension and production develop with age in the two languages of 4–7-year-old Arabic-Swedish-speaking bilinguals without DLD, and how does language exposure and SES influence that development?

RQ2: How do 4–7-year-old Arabic-Swedish-speaking bilinguals without DLD perform on NWR tasks, and how is their performance affected by language exposure, vocabulary size, and properties of the non-words (length, phonological complexity, and language-(non-)specificity)?

RQ3: By comparison, how do Arabic-Swedish-speaking bilingual children with a DLD diagnosis perform on vocabulary and NWR tasks? Does one particular type of NWR task identify DLD better in this bilingual group?

2. Materials and Methods

2.1. Participants

The participants were 110 Arabic-Swedish-speaking children aged 4;0–7;11, 99 with typical language development (the TD sample), and 11 with a diagnosis of DLD (the DLD sample). The two groups will be described in the following.

2.1.1. The TD Sample

A total of 116 children were recruited for the TD sample. They were recruited by contacting a large number of (pre)schools, as well as congregations and associations arranging activities for Arabic-speaking children. Some participants were also recruited via personal contacts of Arabic-speaking members of the research team. Of the 116 children, 17 were excluded for various reasons. The reasons for exclusion included speaking an Arabic variety that was too distant from the prepared dialect versions of the vocabulary tasks (2/17), having only rudimentary knowledge in one language (6/17), not being able to complete all tasks (5/17), not having reached their fourth birthday (2/17), or not speaking one of the target languages (1/17). One child was excluded from the TD sample as it turned out that she was recruited for the DLD sample 18 months later, now having a DLD diagnosis. Only children who could speak both languages were included in the study. The 99 children in the TD sample (see Table 1) attended 53 different (pre)schools in Eastern Central Sweden. According to parental report, the children in the TD sample had no known hearing problems, language disorders or neuropsychiatric disorders at the time of testing.

Table 1. Participants in the TD sample. Number of participants, sex, mean age (years; months) and age range (years; months) per age group.

	4-Year-Olds (n = 22)	**5-Year-Olds (n = 24)**	**6-Year-Olds (n = 29)**	**7-Year-Olds (n = 24)**	**Total (n = 99)**
Girls/boys	12/10	9/15	12/17	16/8	49/50
Mean age	4;5	5;6	6;6	7;7	6;1
Age range	4;0–4;11	5;0–5;11	6;0–6;11	7;0–7;11	4;0–7;11

A bit more than half of the children were born in Sweden (56%), and the rest (42%) had migrated with their families from an Arabic-speaking country (or in one case from a third country). The majority spoke a Levant variety of Arabic (Syrian: 43%, Palestinian: 26%, Lebanese: 9%), 17% spoke Iraqi Arabic, and 4% spoke Egyptian Arabic. Many children were exposed to more than one Arabic variety, beyond the variety spoken in the home. A handful of children were also exposed to a third language in addition to Arabic and Swedish, which was either English, Kurdish (Sorani) or Neo-Aramaic.

In nearly all families, both parents had Arabic as their L1 (96%).[4] In a few cases, information was available for only one parent (3%), or missing for both parents (2%). In one family, one parent stated that their L1 was not Arabic (but presumably Kurdish). Virtually all parents were first-generation immigrants, with residence lengths varying from 10 months to 31 years. A few parents had come to Sweden as children, but most had immigrated as adults. Only one parent had been born in Sweden.

All but one child had received regular input in Arabic from birth. One child was reported to have started to hear Arabic shortly after age 1, and for one child, such information was missing. For Arabic then, there was hardly any variation in age of onset. By contrast, age of onset varied considerably for Swedish. A bit less than half (48%) had an age of onset to Swedish that was before age 3;0 (this included 6% with regular input in Swedish from birth). Twenty children had had less than two years (24 months) of exposure to Swedish at the time of testing. Yet these children were immersed in the Swedish language in preschool and could complete all tasks in Swedish. We decided not to exclude children with short residence lengths or late exposure to Swedish a priori. As long as the children could complete the tasks in both languages, they were included in the study.

All children attended institutional childcare, mostly 25–40 h a week. The 4- and 5-year-olds, as well as four 6-year-olds, attended *förskola* (preschool). All other 6-year-olds attended Swedish-medium *förskoleklass* (a preparatory year for primary school), and the 7-year-olds were in first grade of primary school. Generally, schooling was in Swedish, but 13 children had attended or were attending a bilingual Arabic-Swedish preschool and two children a bilingual English-Swedish preschool, according to parental report.

The children came from a wide variety of socio-economic backgrounds, both concerning parental occupations and education, where all levels from less than six years of primary education to doctorate degrees were represented (i.e., levels 0–8 on the 9-level ISCED 2011 classification, UNESCO Institute for Statistics 2012).

2.1.2. The DLD Sample

Eighteen children with a DLD diagnosis and their parents were invited by their SLP to participate in the study. Of these, four families turned down participation. Three further children had to be excluded, as one child spoke an Arabic variety that was too distant, another child lived too far away for data collection to be feasible, and one child could not be seen due to the outbreak of the COVID-19 pandemic. The 11 children in the final DLD sample were recruited via SLPs working in both public healthcare and private SLP clinics, as well as SLPs working in preschools and schools. Inclusion criteria were: (1) age 4;0–7;11, (2) being regularly exposed to Swedish and an Arabic variety that matched the TD sample (i.e., Levantine, Iraqi or Egyptian), (3) being able to speak at least some Swedish and Arabic, and (4) having a DLD diagnosis. All children had been assessed and diagnosed by a licensed SLP. All but two children (BiAraLI-07 and BiAraLI-09) had been assessed in both Swedish and Arabic by a bilingual SLP or via an interpreter (BiAraLI-08), and (save for one child, BiAraLI-05) had had extensive contact with an SLP in the clinic or at school, often for several years.[5] Diagnoses could include mixed comprehension and production difficulties (Swe: *generell språkstörning*) as well as primarily comprehension difficulties (Swe: *impressiv språkstörning*) or production difficulties (Swe: *expressiv språkstörning*), but not exclusively phonological or articulatory difficulties (Swe: *fonologisk språkstörning*). Exclusion criteria were: (1) having a known biomedical condition associated with language difficulties (e.g., Down syndrome), (2) a diagnosis within the autism spectrum, or (3) intellectual disability. Although not an exclusionary criterion for participating in the DLD study, none of the children had ADHD.

There were fewer girls (4) than boys (7) in the DLD sample, and the age range (5;0–7;3) was narrower compared to the TD sample (4;0–7;11). The mean age of the DLD sample (6;2) was similar to the TD sample (6;1). As can be seen in Table 2, all children were exposed to Arabic from birth, and six of them had received regular exposure to Swedish before age 3. Roughly half of the children (6/11) were reported to have even exposure to both languages, and four of them had slightly more Swedish (60%) than Arabic (40%) in their daily exposure. Only one child was reported to hear mostly Arabic (80%). Six children spoke an Iraqi variety, which differed from the TD sample where only 17% spoke an Iraqi variety. Six children attended preschool, five attended *förskoleklass*, and one child was in first grade of primary school. All children but two had the diagnosis *generell språkstörning* (mixed impressive and expressive language disorder). One child had an unspecified diagnosis, but the SLP suspected *generell språkstörning*, and one child had an expressive language disorder.

Table 2. Age at testing, age of onset for Arabic and Swedish, daily exposure, Arabic variety, (pre)school type, and diagnosis for the children in the DLD sample.

	Age	Age of Onset		Daily Exposure	Arabic Variety	(Pre)school	Diagnosis
		Arabic	Swedish				
BiAraLI-01	6;8	at birth	5;0–6;0	Swe 60%/Ara 40%	Iraqi	Fskklass	Unspec. LD *
BiAraLI-02	6;1	at birth	1;0–2;0	Swe 20%/Ara 80%	Iraqi	Preschool	General LD
BiAraLI-03	5;7	at birth	1;0–2;0	Swe 50%/Ara 50%	Syrian	Preschool	General LD
BiAraLI-04	6;0	at birth	4;0–5;0	Swe 50%/Ara 50%	Syrian	Preschool	General LD
BiAraLI-05	7;3	at birth	4;0–5;0	Swe 60%/Ara 40%	Iraqi	1st grade	Expressive LD
BiAraLI-06	5;4	at birth	1;0–2;0	Swe 50%/Ara 50%	Syrian	Preschool	General LD
BiAraLI-07	6;1	at birth	1;0–2;0	Swe 60%/Ara 40%	Iraqi	Preschool	General LD
BiAraLI-08	7;1	at birth	3;0–4;0	Swe 50%/Ara 50%	Syrian	Fskklass	General LD
BiAraLI-09	6;7	at birth	1;0–2;0	Swe 50%/Ara 50%	Iraqi	Fskklass	General LD
BiAraLI-10	5;0	at birth	1;0–2;0	Swe 50%/Ara 50%	Iraqi	Lang. preschool	General LD
BiAraLI-11	6;4	at birth	4;0–5;0	Swe 60%/Ara 40%	Palestinian	Fskklass	General LD

Note. 'Fskklass' = *förskoleklass*, a preparatory year between preschool and primary school. 'Lang. preschool' = language preschool (*språkförskola*), a specialised preschool unit for children with severe DLD. 'LD' = Language Disorder. * The child had an unspecified diagnosis at the time of testing, but the SLP suspected general LD.

2.2. Materials

2.2.1. Cross-Linguistic Lexical Tasks

The Cross-linguistic lexical task (CLT) is a picture-based vocabulary assessment material (Haman et al. 2015). Each CLT has four subtasks: noun comprehension, verb comprehension, noun production, and verb production. Each part consists of 30 items plus two practice items, making the maximum score 60 for each part, comprehension and production. The comprehension part is a picture selection task. The experimenter asks a prompt question (e.g., 'who is pouring?') and the child has to identify the correct response from an array of four pictures. The production part is a picture-naming task, where the child is shown one picture at a time and is asked to answer the prompt question (e.g., 'what is this?') with a word that corresponds to the picture. The CLT was developed specifically for assessing vocabulary in both languages of bilingual children and is currently available in more than 30 different languages (https://multilada.pl/en/projects/clt/, (accessed 20

June 2022)). For a detailed description of the construction of the CLT, please see Haman et al. (2015). In the current study, the Swedish version (Ringblom et al. 2014) and an Arabic CLT version (Haddad 2017) that was adapted from the Lebanese Arabic version (Khoury Aouad Saliby et al. 2017a) were used. Since only a few of the children in the present study spoke Lebanese Arabic, the existing Lebanese version was adapted to the Arabic varieties most relevant to the Swedish context. For the CLT comprehension tasks, new prompts were constructed for all test items in the respective dialect, so that no child was disadvantaged by being asked about a word in a dialect they were not familiar with. For the CLT production tasks, the Lebanese target words needed to be complemented by other dialect synonyms, particularly Syrian, Palestinian and Iraqi, as well as Modern Standard Arabic (MSA). Four different adaptations were developed for Syrian, Palestinian, Lebanese and Iraqi Arabic (Haddad 2017).[6]

2.2.2. Non-Word Repetition Tasks

In the present study, three NWR tasks were used, all developed for children of preschool and early school age. First, a Swedish language-specific task (LS-Swe), originally developed by Barthelom and Åkesson (1995), was used, for which reference data for 4–6-year-old monolinguals is available (Radeborg et al. 2006). The LS-Swe encompasses 24 test items of 2–5 syllables (6 of each syllable length) that adhere to Swedish phonotactics and contain phonemes that are typical of Swedish; nineteen consonant phonemes (/p, b, t, d, k, ɡ, m, n, ŋ, ɾ, f, v, s, ɕ, ɧ, ʂ, h, j, l/) and fifteen vowel phonemes (/i, ɪ, y, ʏ, e, ɛ, œ, ɑ, a, o, ɔ, u, ʊ, ʉ, ɵ/). The items have syllables with varying phonological complexity: there are open and closed syllables with and without consonant clusters (13 items with no clusters, 9 items with one cluster, and 2 items with two clusters) in onset and coda. The items are pronounced with stress patterns that are typical of Swedish, i.e., with varying main stress and vowel duration in different syllables, for example /spɵrɪfɾaˈɡoːl/ and /flɛtɛˌmɪŋɛˈɾoːf/. The LS-Swe items were recorded by a female speaker of Swedish speaking a central Swedish dialect, which is close to standard Swedish.

Second, a Swedish version of the cross-linguistic NWR task (Chiat 2015) was used (CL-Swe). The task was designed to be compatible with the lexical phonology of many languages. As such, it contains items of 2–5 syllables (4 of each syllable length), with no consonant clusters and no codas (only open syllables). The full range of phonemes includes eleven consonants (/p, b, t, d, k, ɡ, s, z, m, n, l/) and three vowels (/a, i, u/). For the purpose of this study, a Swedish version was created. From a list of 84 candidate items, 16 items were chosen, for example, /lɪmɪka/and/tʊlɪɡasʊmʊ/, excluding items that contain phonemes that do not exist in Swedish (e.g., /z/), or contain real words or inflections in that language. The CL-Swe items were recorded by the same female speaker who recorded the LS-Swe items. All items were pronounced with quasi-neutral prosody (Chiat 2015, p. 138), where all syllables were equally stressed (i.e., they carried equal length and pitch) apart from final-syllable lengthening and pitch drop marking the end of an utterance.

Finally, the third task was the Non-word repetition task-Lebanese (NWRT-Leb, Abou Melhem et al. 2011), a Lebanese version of the QU task, modelled on the NWR-FRENCH task (dos Santos and Ferré 2018). This task was constructed to investigate how phonological complexity impacts NWR performance. The task contains 30 items of 1–3 syllables (6 items with one syllable, 14 items with two syllables and 10 items with three syllables) with and without consonant clusters (15 items with no clusters, 13 items with one cluster, and 2 items with two clusters) and codas. There were three different types of syllables, all present in Lebanese Arabic (and also in other spoken varieties of Arabic), French and English: CV, CCV or CVC. This task includes only seven phonemes, four consonants (/b, l, k, f/) and three vowels (/a, i, u/), phonemes that all exist in Lebanese (and other varieties of) Arabic, French and English. Each item contained three to seven phonemes, for instance /fablu/ and /bifakub/. The NWRT-Leb items were recorded at the Department of Speech and Language Therapy, St Joseph University, by a female speaker of Lebanese Arabic.

For all three NWR tasks, audio files were created where each item was played one after the other, with a three second pause in between each non-word. The LS-Swe and the CL-Swe were presented with increasing level of difficulty, i.e., starting with the items that were the shortest (had the lowest number of syllables), and gradually increased with one additional syllable. The NWRT-Leb items were presented in randomised order. The audio recordings were incorporated into audio-visual PowerPoint presentations. A list of all items in the NWR tasks is provided in Table A1 in the Appendix A.

2.2.3. Parental Questionnaire

The parental questionnaire used in the present study was developed for a large-scale childhood multilingualism research project at Uppsala University, BiLI-TAS (PI: Ute Bohnacker; BiLI-TAS is short for Bilingualism & Language Impairment Turkish/Arabic/Swedish). The questionnaire could be answered in either Arabic or Swedish, whichever the parents preferred. The questionnaire provided information about the social and linguistic background of the children and their parents. The questions targeted (early) language development, family history of speech, language and literacy difficulties, concerns about language development, language exposure, language use in the family, language activities in the home such as book reading and storytelling, as well as parental education, occupation and language skills.

The questionnaire administered to the parents of the children in the DLD sample was identical to the TD questionnaire, but in addition included questions that queried for how long the child had been in contact with an SLP, and who took the initiative for SLP assessment (e.g., parents, preschool staff, or the child healthcare nurse).

2.2.4. Interviews with Parents, Teachers and SLPs of the DLD Children

The questions for the interviews with parents, SLPs and teachers were developed by the BiLI-TAS team, and first used during a clinical study of Turkish-speaking children, as described in Öztekin (2019). The original interview templates were slightly modified in order to suit the current study. The parents were interviewed in connection with Arabic data collection in the home, or by telephone. The questions asked during the interview concerned the same topics as in the questionnaire, but provided more in-depth information, for instance on how the parents viewed their child's language development over time, their attitudes and beliefs regarding language development and bilingualism, and whether they were concerned about their child's language development. The teachers were interviewed during a preschool visit, in connection with data collection in Swedish. Teacher interviews included questions about the child's language skills, their communicative and social behaviour, whether they could follow instructions, and how they behaved during book reading and group activities that promoted linguistic awareness (e.g., rhyming and language games). The SLPs were interviewed by telephone. Interview questions included: how the child had been assessed (in which language(s), and which materials were used), age at referral, language therapy and the child's development over time, the parents' attitudes towards therapy, current diagnosis, and what the SLP considered to be most striking or problematic about the child's language.

2.3. Procedure

The study was planned and carried out in accordance with Swedish legislation on research ethics and data protection, and adheres to the university ethical code of conduct (Codex) that came into place halfway through the BiLI-TAS research project. Prior to participation, the parents of all children gave their informed written consent. They could revoke their participation at any time.

Data Collection

Data were collected between September 2017 and March 2019 for the TD sample, and between January and September 2019 for the DLD sample.

The children were assessed with the CLT and NWR tasks as part of a test battery that also included narrative tasks and a fourth NWR task (an Arabic version of the CL task).[7] Each child was seen on two separate occasions, one in each language, either at (pre)school, in the home, or at a community centre, with each session lasting 30–45 min. The median interval between sessions was 7 days. The order of the languages as well as the order of the tasks were counterbalanced. Tasks were administered by trained native speakers, and the experimenter spoke to the child only in the language of testing in order to be able to assess the children's knowledge of each language separately. The dialectal CLT items and the Arabic variety spoken by the experimenter were matched with the variety spoken by the child. Sessions were video- and audio-recorded, so that all responses could be checked afterwards.

The CLT was administered via coloured picture booklets, following the standard procedure described by Haman et al. (2015). During the session, responses were noted on paper forms, and the experimenter gave only neutral feedback (e.g., *aha*, *mhm*, *okay*) irrespective of whether the child had provided a correct answer or not. After each session, responses were transcribed and scored.

The NWR tasks were administered via audio-visual PowerPoint presentations and presented to the children as an imitation task. The LS-Swe and the CL-Swe tasks feature a parrot and the NWRT-Leb features an alien that the child is instructed to imitate. The task was presented to the child on a smartphone, and the audio was played via noise-cancelling headphones. All tasks were audio-recorded, and the responses were transcribed and scored after the session.

The child was always praised at the end of each task, irrespective of the actual outcome, and rewarded with stickers.

2.4. Data Treatment

2.4.1. Scoring

All CLT child responses were transcribed and scored. The maximum score for each subtask was 60 points. Every child completed all four subtasks. The total number of responses was 26,400 (= 110 children × 2 languages × 120 test items (i.e., 60 for comprehension + 60 for production)). The scoring was done by native speakers of Swedish and Arabic. As there is no standardised published procedure for scoring the CLTs, scoring was done as follows. One point was awarded for each correct response in the language of testing. For the comprehension tasks, only target picture identification was scored as correct. For the production tasks, a point was awarded if the child produced the target word, for example, Arabic *dabdab, zaḥaf* or *ḥaba* ('crawl') on the Arabic CLT, or Swedish *krypa* ('crawl') on the Swedish CLT, in response to a picture of a baby crawling. Moreover, the following responses were also scored as correct: (i) adult-like synonyms, (ii) words that were more specific than the target word and corresponded to the picture (e.g., Swe. *meta* 'to angle' instead of the target *fiska* 'to fish'), and (iii) word forms that were pronounced slightly off-target but were still recognisable as the target lemma. All other types of responses were scored as incorrect. Thus, words not in the target language, words that corresponded to the picture but were less specific than the target word (e.g., Swe. *städa* 'clean' instead of *sopa* 'sweep'), paraphrases and circumlocutions, forms belonging to a different word class, and forms that phonologically and/or morphologically strongly deviated from the target word, were scored zero. The scoring of items was carefully checked for consistency. Unclear items were discussed by the authors and Arabic- and Swedish-speaking team members until consensus was reached. Whenever necessary, the audio and video recordings were consulted.

All NWR tasks were audio-recorded for later transcription and analysis. The total number of responses was 7636 (2592 for LS-Swe (108 participants × 24 test items), 1744 for CL-Swe (109 participants × 16 test items), and 3300 for NWRT-Leb (110 participants × 30 test items)). The responses were transcribed phonemically by a native speaker of Swedish (LS-Swe and CL-Swe) and Arabic (NWRT-Leb), respectively. As there is no standardised procedure for scoring any of these NWR tasks, scoring was done as follows. The participants received 1 point for each correctly repeated non-word, and 0 points for any response containing an error (the whole item correct vs. incorrect approach). Allowances were made for minor articulation deviances, such as non-adultlike or indistinct pronunciation of /r/ and /s/ (as these phonemes are challenging to articulate and may be difficult to pronounce even for some adults). Any phonological substitution processes that were consistent in the child's speech were disregarded. Errors of voicing (/p/ vs. /b/) and minor vowel deviations (e.g., /oe/ vs. /ø/) were also disregarded. However, major vowel substitutions, such as substituting /a/ for /i/, were not allowed. Finally, any additions of syllables or phonemes before or after the otherwise correctly repeated item were also disregarded (i.e., children were not penalised for hesitation noises). The scoring of items was carefully checked for internal consistency. Moreover, for interrater reliability, an independent researcher transcribed and scored the responses of 15 randomly sampled participants, 12 TD children (12%) and three DLD children (27%), for all three NWR tasks. The interrater agreement rate was 98.0% (1029/1050 items).

2.4.2. Questionnaire Data

In the present study, four variables from the questionnaire data were investigated with respect to the performance on the vocabulary and the NWR tasks: chronological age, length of exposure, daily exposure, and SES. In the following, it will be described how they were operationalised.

First, chronological age was the child's age at testing, measured in number of months. Second, age of onset (AoO) for Swedish was the reported age at which the child started to receive regular exposure to Swedish. AoO was transformed into Length of Exposure to Swedish (LoESwe) by subtracting AoO (months) from the child's chronological age (months). As AoO for Arabic was at birth for all but two children,[8] there was an almost complete overlap between chronological age and LoE for Arabic, and they could not be investigated as separate variables. Third, the child's current daily exposure to each language was estimated by the parents on a scale with seven levels ranging from almost only Arabic (95% Arabic and 5% Swedish) to almost only Swedish (5% Arabic and 95% Swedish). Parents could also note a different distribution. For the purpose of statistical analyses, the variable of daily exposure was split into two separate variables, one for each language. Daily exposure to Arabic (Daily exp Ara) thus indicated the percentage of daily exposure to Arabic, and Daily exposure to Swedish (Daily exp Swe) indicated the percentage of daily exposure to Swedish. Finally, SES was operationalised as parental education. The questionnaire queried the highest level of education of each parent. The responses were coded according to the 9-level ISCED 2011 classification of education (UNESCO Institute for Statistics 2012). Then, the education level was averaged across both parents. For a couple of children, information was available only for one parent (e.g., single-parent households). In such cases, the SES variable was based on the education level of that one parent.

2.4.3. Interview Data

The interview data was arranged thematically in a spreadsheet according to the questions posed to the informant. Next, all responses were systematically searched for descriptions of the child's language abilities, as well as their communicative behaviour. The parents' answers were further searched for information about delayed language development, and the teacher's answers were searched for information on the child's classroom behaviour and peer relations. Finally, the SLPs answers were searched for descriptions of behaviour and progress in assessment and therapy. For a condensed overview, see Table A2 in the Appendix A. More details are provided in Öberg (2020).

2.4.4. Statistical Analyses

All analyses were conducted in R (R Core Team 2021). Questionnaire data was missing for one seven-year-old in the TD sample, so this participant was excluded from all analyses that contained background variables. All correlations were calculated with Pearson's correlation coefficient (Pearson's r). For all statistical analyses, the level of significance was set at $p < 0.05$ (two-tailed).

For all vocabulary and NWR tasks, age development was investigated by correlating age in months with raw scores. Vocabulary comprehension and production were investigated separately for each language. Due to the different number of items (and thus, different maximum scores) in the three NWR tasks, total scores were converted into proportions before the performance on the three tasks was compared with a one-way ANOVA. Bonferroni correction for multiple comparisons was used in the subsequent post-hoc tests.

For vocabulary, multivariate linear regression models were fitted, with vocabulary score as the dependent variable. Comprehension and production were analysed separately for each language, thus there were four separate models. All independent variables were centred before modelling; thus, the intercept indicates the mean of the whole sample. As SES (parental education) data was missing for five children, SES data was imputed using regression imputation in order to avoid excluded data points in the sample.

Item-related and participant-related effects on the accuracy of repetition of NWR items were investigated with logistic mixed-effects regression models, using the function *glmer* from the *lme4* package (Bates et al. 2015). The dependent variable, *accuracy*, was a categorical variable, where each data point indicates correct (1p) or incorrect (0p) repetition of a NWR item. All continuous variables were standardised prior to modelling. Mixed-effects regression models account for dependencies in the data by so-called random effects. This type of model is suitable when data points are not independent of each other. Since participants and non-word items are repeated many times in the data set, random effects account for these dependency structures. In sum, all of the logistic mixed-effects models investigated which of the independent variables could predict whether a response was correct or not, while accounting for non-independence. The mixed-effects models were evaluated with pseudo-R^2 and concordance index (c-index). Pseudo-R^2 was obtained with the *r.squaredGLMM* function from the *MuMIn* package (Bartoń 2020). For mixed-effects models, the marginal R^2 expresses the amount of variance that is explained by the fixed effects alone, and the conditional R^2 expresses the amount of variance explained by the full model, including random effects (Nakagawa et al. 2017; Nakagawa and Schielzeth 2013). The c-index is a measure of concordance between a model's predicted probabilities for each data point and the actual outcome. A value above 0.8 is generally considered to be a good model (Baayen 2008, p. 204).

In order to compare the performance of the children in the DLD sample on the vocabulary and NWR tasks to that of the TD sample, age-adjusted z-scores were calculated. First, z-scores were calculated for all children in the TD sample, based on the raw score for each child and the mean and SD for that child's age group. Next, the raw scores of the children in the DLD group were transformed into z-scores, based on the mean and SD for the corresponding age group in the TD sample. Thus, all z-scores indicate how each individual performed on a specific task compared to age-group peers in the TD sample.

3. Results

3.1. Vocabulary in the TD Sample

3.1.1. Descriptive Statistics

First, total scores and scores by age groups are reported separately for comprehension and production in Arabic and Swedish. As can be seen in Table 3, scores increased with age in all tasks.[9]

Table 3. Means, standard deviations (SD), and ranges for each CLT vocabulary task by age groups and total scores. Maximum score for all tasks = 60 points.

		4-Year-Olds (n = 22)	**5-Year-Olds (n = 24)**	**6-Year-Olds (n = 29)**	**7-Year-Olds (n = 24)**	**Total (n = 99)**
Arabic comprehension	Mean (SD) Range	41.5 (7.6) 25–52	46.7 (6.6) 27–56	48.5 (7.5) 31–58	52.4 (3.6) 45–59	47.5 (7.5) 25–59
Arabic production	Mean (SD) Range	25.5 (12.2) 1–42	32.7 (12.0) 11–48	34.5 (13.2) 10–53	37.1 (9.2) 16–52	32.7 (12.3) 1–53
Swedish comprehension	Mean (SD) Range	36.0 (8.3) 18–52	45.2 (9.0) 29–59	46.9 (10.2) 27–60	53.3 (8.5) 27–60	45.7 (10.8) 18–60
Swedish production	Mean (SD) Range	22.3 (7.2) 10–41	29.7 (9.8) 15–48	31.6 (11.5) 11–48	39.5 (11.2) 12–53	31.0 (11.7) 10–53

In the following, age will be treated as a continuous variable. There were positive correlations between linear age and scores on all vocabulary tasks. The correlation with age was stronger for comprehension than production in both languages (Arabic comprehension: $df = 97$, $r = 0.50$, $p < 0.001$; Arabic production: $df = 97$, $r = 0.33$, $p < 0.001$; Swedish comprehension: $df = 97$, $r = 0.51$, $p < 0.001$; Swedish production: $df = 97$, $r = 0.46$, $p < 0.001$).

3.1.2. Age, Language Exposure, SES and Vocabulary

In order to investigate the relative effect of age (in months), amount of daily exposure (to Arabic or Swedish), length of exposure (for Swedish) and SES on vocabulary scores, four multivariate linear regression models were run, separately for comprehension and production in Arabic and Swedish, respectively (see Tables 4 and 5). Model 1 explained 36% of the variance for Arabic comprehension scores, with only age and daily exposure being significant predictors. As is evident from the standardised estimates, age was a stronger predictor ($\beta = 0.57$, $p < 0.001$) than daily exposure ($\beta = 0.37$, $p < 0.001$). Model 2 explained 36% of the variance for Arabic production scores. For both Arabic production and comprehension, only age and daily exposure to Arabic were significant, but for Arabic production, daily exposure ($\beta = 0.52$, $p < 0.001$) was a stronger predictor than age ($\beta = 0.41$, $p < 0.001$). SES was not a significant predictor of Arabic comprehension ($p = 0.55$) or production ($p = 0.19$) scores.

Table 4. Multivariate linear regression models for Arabic comprehension and Arabic production. Estimates (B), standard errors (SE (B)), t-scores (t), *p*-values (*p*) and standardised estimates (β).

	Model 1 (Arabic Comprehension)					**Model 2 (Arabic Production)**				
	B	**SE (B)**	**t**	***p***	**β**	**B**	**SE (B)**	**t**	***p***	**β**
Intercept	47.47 ***	0.61	78.18	<0.001		32.60 ***	1.00	32.65	<0.001	
Age	0.31 ***	0.04	6.91	<0.001	0.57	0.36 ***	0.07	4.93	<0.001	0.41
Daily exp Ara	0.18 ***	0.04	4.50	<0.001	0.37	0.42 ***	0.07	6.26	<0.001	0.52
SES	0.21	0.35	0.59	0.55	0.05	−0.76	0.58	−1.32	0.19	−0.11
	R^2 (adj.) = 0.36, $F(3,94) = 19.58$, $p < 0.001$					R^2 (adj.) = 0.36, $F(3,94) = 19.36$, $p < 0.001$				

Note. VIF values were around 1 for all predictors, indicating low levels of collinearity. *** $p < 0.001$.

Model 3 explained 53% of the variance for Swedish comprehension scores. Only age, length of exposure and daily exposure were significant predictors, with length of exposure ($\beta = 0.42$, $p < 0.001$) being the most influential, followed by age ($\beta = 0.35$, $p < 0.001$) and daily exposure ($\beta = 0.17$, $p < 0.05$). Similar patterns were found for Swedish production, where Model 4 explained 51% of the variance. Again, only age, length of exposure and daily exposure were significant predictors, with length of exposure ($\beta = 0.43$, $p < 0.001$) having more of an impact than age ($\beta = 0.28$, $p < 0.001$) or daily exposure ($\beta = 0.24$, $p < 0.01$).

SES was not a significant predictor of Swedish comprehension (p = 0.09) or production (p = 0.33) scores.

Table 5. Multivariate linear regression models for Swedish comprehension and Swedish production. Estimates (B), standard errors (SE (B)), t-scores (t), p-values (p) and standardised estimates (β).

	Model 3 (Swedish Comprehension)					**Model 4 (Swedish Production)**				
	B	**SE (B)**	**t**	***p***	**β**	**B**	**SE (B)**	**t**	***p***	**β**
Intercept	45.85 ***	0.74	61.90	<0.001		31.18 ***	0.82	37.97	<0.001	
Age	0.26 ***	0.06	4.44	<0.001	0.35	0.23 ***	0.07	3.46	<0.001	0.28
LoE Swe	0.23 ***	0.05	5.09	<0.001	0.42	0.26 ***	0.05	5.13	<0.001	0.43
Daily exp Swe	0.12 *	0.05	2.37	0.02	0.17	0.18 **	0.06	3.19	0.002	0.24
SES	0.75	0.44	1.72	0.09	0.12	0.47	0.49	0.97	0.33	0.07
	R^2 (adj.) = 0.53, F(4,93) = 28.1, p < 0.001					R^2 (adj.) = 0.51, F(4,93) = 25.78, p < 0.001				

Note. VIF values were around 1 for all predictors, indicating low levels of collinearity. *** p < 0.001, ** p < 0.01, * p < 0.05.

3.2. NWR in the TD Sample

3.2.1. Descriptive Statistics

First, total scores and scores by age groups are reported for all three NWR tasks. As can be seen in Table 6, overall performance was lowest on the LS-Swe task, with a mean accuracy of 55.0%. The CL-Swe task was in the middle, with a mean accuracy of 76.1%, and the NWRT-Leb task had the highest overall performance, with a mean accuracy of 83.7%. A one-way ANOVA revealed that the differences in accuracy between tasks were significant (F(2, 291) = 97.65, p < 0.001, $\eta_p{}^2$ = 0.40). Pairwise comparisons showed that the differences were significant between all tasks (p < 0.01 for all comparisons).

As shown in Table 6, mean scores increase with age between all age groups, and the ranges generally decrease.[10] In the following analyses, age will be treated as a continuous variable. There were positive correlations between linear age and scores on all NWR tasks, but they were slightly weaker for the CL-Swe task (df = 96, r = 0.27, p < 0.01) than for the LS-Swe task (df = 95, r = 0.41, p < 0.001) and for the NWRT-Leb (df = 97, r = 0.45, p < 0.001).

Table 6. Mean scores, standard deviations (SD), ranges, mean accuracies in %, and SDs for each NWR task by age groups and total scores.

		4-Year-Olds (n = 22)	**5-Year-Olds (n = 24)**	**6-Year-Olds (n = 29)**	**7-Year-Olds (n = 24)**	**Total (n = 99)**
LS-Swe (Max = 24)	Mean (SD)	10.6 (3.6)	12.4 (3.6)	14.1 (3.9)	15.0 (3.4)	13.2 (4.0)
	Range	3–16	6–18	6–21	8–22	3–22
	Mean (SD) %	44.2 (15.0)	51.7 (15.0)	58.9 (16.4)	62.7 (14.3)	55.0 (16.5)
CL-Swe (Max = 16)	Mean (SD)	11.0 (2.4)	12.3 (2.0)	12.3 (2.4)	12.9 (1.8)	12.2 (2.2)
	Range	5–14	8–15	8–16	10–16	5–16
	Mean (SD) %	69.0 (15.1)	76.8 (12.4)	76.7 (15.0)	80.7 (11.0)	76.1 (13.9)
NWRT-Leb (Max = 30)	Mean (SD)	21.3 (4.9)	25.3 (3.8)	26.3 (3.4)	27.0 (2.6)	25.1 (4.2)
	Range	7–28	15–30	19–30	19–30	7–30
	Mean (SD) %	71.1 (16.3)	84.4 (12.6)	87.7 (11.3)	89.9 (8.6)	83.7 (14.1)

Note. For LS-Swe, N = 97, as only 20/22 four-year-olds did the task. For CL-Swe, N = 98, as only 21/22 four-year-olds completed the task.

There were differences in the proportion of children who scored high or low on each task, reflecting different overall task difficulty. For instance, there was a striking difference between the LS-Swe task and the NWRT-Leb, where 38% of all children scored below 50% on the LS-Swe task, but only 3% did so on the NWRT-Leb. The reverse pattern emerged when investigating the proportion of children who scored 90% or better; only one child did so on the LS-Swe task, but 41% did so on the NWRT-Leb.[11] For the CL-Swe task, most children scored between 50–90% correct, with fewer children scoring below 50% (10% of the children) or above 90% (12% of the children).

3.2.2. NWR Accuracy in Relation to Task, Item Properties, Language Exposure, and Vocabulary

As described in the literature (see Introduction), several item-/task-related factors and participant-related factors have an impact on NWR performance. First, exploratory analyses were conducted for NWR tasks by investigating correlations with language exposure and vocabulary. Unsurprisingly, for LS-Swe, there were positive correlations with length of exposure to Swedish ($df = 94$, $r = 0.36$, $p < 0.001$) and Swedish vocabulary comprehension ($df = 95$, $r = 0.45$, $p < 0.001$). More surprisingly, for the CL-Swe, there were also positive correlations with length of exposure to Swedish ($df = 95$, $r = 0.23$, $p < 0.05$) and Swedish vocabulary comprehension ($df = 96$, $r = 0.37$, $p < 0.001$). For the NWRT-Leb, there was a positive correlation with Arabic vocabulary comprehension ($df = 97$, $r = 0.36$, $p < 0.001$). For none of the NWR tasks were there any correlations with daily exposure or SES.

Next, accuracy in terms of the percent of correctly repeated items was investigated in relation to item length (number of syllables), as visualised separately for each NWR task in Figure 1.

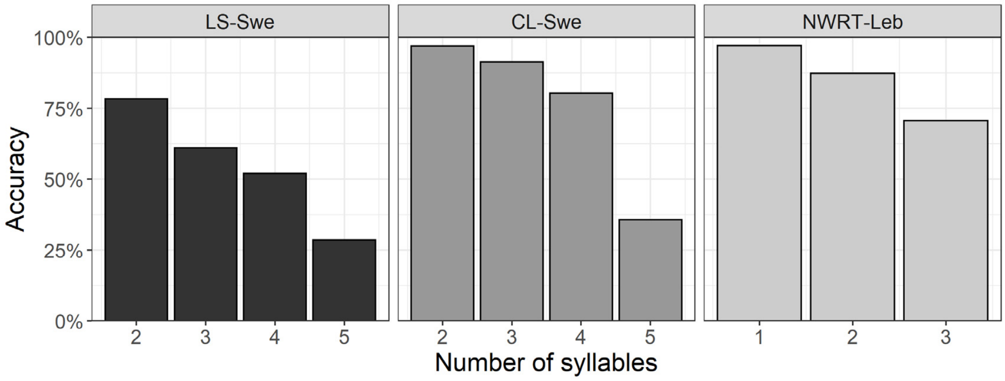

Figure 1. Accuracy (% correct responses), LS-Swe, CL-Swe and NWRT-Leb by number of syllables.

As presented in Table 7, accuracy was generally highest for items with fewer syllables and without consonant clusters. In other words, accuracy decreased as a function of increasing number of syllables and the presence of consonant clusters. However, the accuracy patterns for items with vs. without clusters were not the same for the LS-Swe task and the NWRT-Leb. In the NWRT-Leb, accuracy decreased by similar amounts for items with and without clusters as the number of syllables increased. By contrast, for the LS-Swe task, accuracy levels were similar for 2–4-syllable items without clusters, but decreased steeply at five syllables, whilst accuracy levels for the items with clusters decreased for each added syllable. This difference is visualised in Figure 2. Potential interactions between item length and presence of clusters will be explored further in the multivariate mixed-effects regression models.

Table 7. Accuracy (% correct responses) by task and number of syllables for all items, items with clusters and items without (*w*/*o*) clusters.

	LS-Swe			**CL-Swe**	**NWRT-Leb**		
	All Items	*w/o Clusters*	*With Clusters*	*All Items*	*All Items*	*w/o Clusters*	*With Clusters*
1 syllable	–	–	–	–	97%	97%	97%
2 syllables	78%	77%	82%	97%	87%	92%	82%
3 syllables	61%	78%	44%	91%	71%	73%	68%
4 syllables	52%	80%	24%	80%	–	–	–
5 syllables	29%	44%	14%	36%	–	–	–

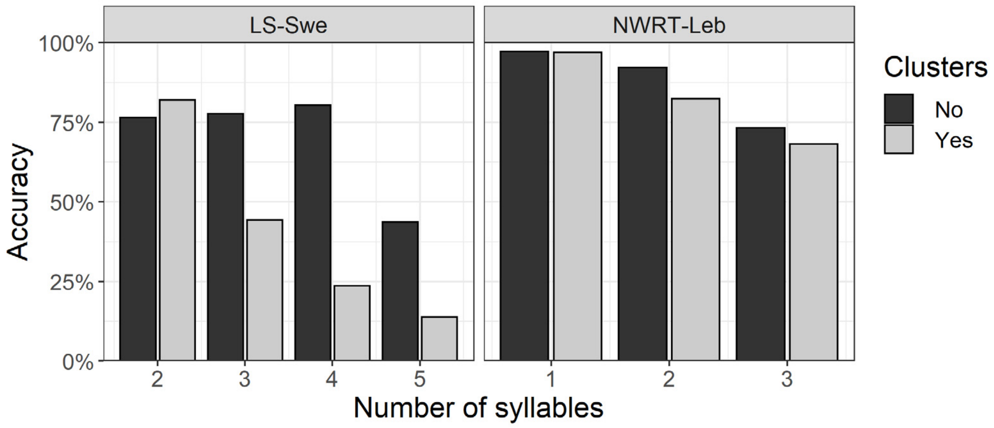

Figure 2. Accuracy (% correct responses) for items in LS-Swe and NWRT-Leb by number of syllables and presence of consonant clusters.

Next, logistic mixed-effects regression models were fitted to investigate the effect of task-/item-related factors and participant-related factors on repetition accuracy. Since the non-word tasks differed from each other on a number of fundamental properties (i.e., number of syllables, presence of consonant clusters, language-(non-)specificity, etc.), all tasks could not be directly compared to each other. Therefore, three separate models were fitted.

Model 5 (see Table 8) investigated the effect of the presence of clusters, item length (number of syllables), age, Arabic vocabulary, and the interaction between presence of clusters and syllables on the repetition accuracy in the NWRT-Leb. Chronological age (B = 0.47, p = 0.001) and Arabic vocabulary (B = 0.32, p = 0.02) had a positive influence on NWR scores. That is, older children and children with larger vocabularies in Arabic had higher accuracy. Consonant clusters (B = −0.72, p = 0.02) and an increasing number of syllables (B = −1.13, p < 0.001) both contributed to lower repetition accuracy. However, there was no interaction between clusters and syllables (B = 0.05, p = 0.88), demonstrating that as the number of syllables increased, accuracy levels decreased alike for items with and without clusters. The full model's explanatory power was considerable (conditional R^2 = 0.52) and larger than that of the fixed effects alone (marginal R^2 = 0.27). The c-index of 0.88 indicates a good model fit.

Table 8. Summary of Model 5: accuracy, NWRT-Leb task.

Random Effects	Variance	SD	*Fixed Effects*	B	SE (B)	z	*p*
Participant	1.13	1.06	Intercept	2.91 ***	0.26	11.11	<0.001
Item	0.60	0.78	Age	0.47 **	0.14	3.27	0.001
			Arabic vocabulary	0.32 *	0.14	2.25	0.02
			Clusters (no vs. yes)	−0.72 *	0.32	−2.25	0.02
			Syllables	−1.13 ***	0.24	−4.76	<0.001
			Clusters (no) × syllables	0.05	0.33	0.15	0.88
Model evaluation	**Marginal R^2** 0.27		**Conditional R^2** 0.52			**C-index** 0.88	

Note. Logistic mixed-effects regression model with random effects: random intercepts for participant and test item. Model fit with maximum likelihood (Laplace approximation). The reference level for categorical variables is the first category. *** p < 0.001, ** p < 0.01, * p < 0.05.

Model 6 (see Table 9) explored the effect of age, length of exposure to Swedish, Swedish vocabulary, clusters, item length (number of syllables), and the interaction between clusters and number of syllables on LS-Swe accuracy. While chronological age (B = 0.29, p = 0.01) had a positive effect on the repetition accuracy of LS-Swe items, there was no effect of length of exposure to Swedish (B = 0.10, p = 0.43). Swedish vocabulary scores (somewhat surprisingly) had no effect (B = 0.26, p = 0.06) on LS-Swe accuracy.[12] The presence of consonant clusters had a negative impact (B = −1.69, p < 0.001), and accuracy also decreased with increasing item length (syllables: B = −0.65, p = 0.007). Furthermore, there was an interaction between clusters and syllables (B = −0.81, p = 0.03), where the negative effect of clusters was stronger with increasing item length (number of syllables). The explanatory power of the full model was considerable (conditional R^2 = 0.53) and greater than that of the fixed effects alone (marginal R^2 = 0.34). The c-index was 0.88, indicating a good model fit.

Table 9. Summary of Model 6: accuracy, LS-Swe task.

Random Effects	Variance	SD	*Fixed Effects*	B	SE (B)	z	*p*
Participant	0.62	0.79	Intercept	1.16 ***	0.26	4.42	<0.001
Item	0.72	0.85	Age	0.29 *	0.12	2.46	0.01
			LoE Swedish	0.10	0.13	0.80	0.43
			Swedish vocabulary	0.26	0.14	1.88	0.06
			Clusters (no vs. yes)	−1.69 ***	0.37	−4.59	<0.001
			Syllables	−0.65 **	0.24	−2.70	0.007
			Clusters (no) × syllables	−0.81 *	0.37	−2.16	0.03
Model evaluation	**Marginal R^2** 0.34		**Conditional R^2** 0.53			**C-index** 0.88	

Note. Logistic mixed-effects regression model with random effects: random intercepts for participant and test item. Model fit with maximum likelihood (Laplace approximation). The reference level for categorical variables is the first category. *** p < 0.001, ** p < 0.01, * p < 0.05.

Finally, Model 7 (see Table 10) investigated the effect of task for the LS-Swe and the CL-Swe (a comparison between language-specific vs. non-language-specific test items), age, Swedish vocabulary, item length, as well as the interaction between task and item length and the interaction between task and Swedish vocabulary scores. Chronological age (B = 0.27, p = 0.01) and Swedish vocabulary scores (B = 0.35, p = 0.002) had a positive effect on repetition accuracy. There was a task effect (B = 1.76, p < 0.001); items from the CL-Swe task generally had higher accuracy than items from the LS-Swe task. Accuracy decreased with increasing item length (syllables: B = −1.08, p < 0.001), and there was also an interaction between task and item length (B = −0.80, p = 0.04). The adverse effect of item length was not the same for both tasks. Accuracy decreased by similar amounts for each added syllable in the LS-Swe task (see Figure 1 and Table 7). By contrast, for the CL-Swe task, the decrease was rather small for each added syllable in the shortest items, and then dropped steeply at five syllables. Finally, there was no interaction between task and Swedish vocabulary (B = 0.04, p = 0.70), indicating that the positive effect of Swedish vocabulary was similar for the language-specific items in LS-Swe and the non-language-specific items in CL-Swe. The explanatory power of the full model was considerable (conditional R^2 = 0.61), and better than that of the fixed effects alone (marginal R^2 = 0.38). The c-index of 0.90 showed a good model fit.

Table 10. Summary of Model 7: accuracy, LS-Swe vs. CL-Swe task.

Random Effects	Variance	SD	r^2	*Fixed Effects*	B	SE (B)	z	*p*
Participant	0.63	0.80		Intercept	0.32	0.24	1.32	0.19
Task LS-Swe (slope)	0.19	0.44	0.25	Age	0.27 *	0.11	2.46	0.01
Item	1.20	1.09		Swedish vocabulary	0.35 **	0.11	3.11	0.002
				Task (LS-Swe vs. CL-Swe)	1.76 ***	0.38	4.64	<0.001
				Syllables	−1.08 ***	0.23	−4.64	<0.001
				Task (LS-Swe) × syllables	−0.80 *	0.38	−2.11	0.04
				Task (LS-Swe) × Swe vocab	0.04	0.11	0.39	0.70
Model evaluation	**Marginal R^2** 0.38			**Conditional R^2** 0.61			**C-index** 0.90	

Note. Logistic mixed-effects regression model with random effects: random intercepts for participant and test item, and by-participant random slopes for task. Model fit with maximum likelihood (Laplace approximation). The reference level for categorical variables is the first category. *** p < 0.001, ** p < 0.01, * p < 0.05.

3.3. The DLD Sample

3.3.1. Performance on Vocabulary and NWR Tasks

In this section, the performance of the children in the DLD sample will be compared to that of the TD group. As the NWR tasks and the CLT vocabulary tasks utilised in the present study have not been normed, we had no indication which cut-off would be best to identify DLD in our sample of Arabic-Swedish speaking bilinguals on these particular tasks. Therefore, we opted for a cut-off of z-score below −1.25 (i.e., identifying the lowest-scoring 10.6%) in accordance with Tomblin et al. (1997).

Most children in the DLD group scored within the range of their age group in both vocabulary comprehension and production in both languages, but below the mean, which is shown by the predominantly negative z-scores (see Table 11). For the children in the DLD group as a whole, vocabulary z-scores were generally lower in Arabic than in Swedish. The low vocabulary scores in Arabic are noteworthy, since the DLD children (like the TD group) had ample exposure to Arabic from birth. For NWR, a bit more than half (6/11) of the children in the DLD sample scored within the range of their age peers on all three NWR tasks. Overall performance on the NWR tasks was generally below the mean, as reflected by the overall negative z-scores.

Table 11. Vocabulary comprehension (comp), vocabulary production (prod), and NWR scores (raw scores and age-adjusted z-scores) for children in the DLD sample.

	Age	Score	Arabic		Swedish		LS-Swe	CL-Swe	NWRT-Leb
			Comp	Prod	Comp	Prod			
BiAraLI-01	6;8	raw score	37	18	29	13	9	7	18
		z-score	**−1.53***	−1.25	**−1.76***	**−1.62***	**−1.31***	**−2.19***	**−2.45***
BiAraLI-02	6;1	raw score	38	25	36	24	5	8	17
		z-score	**−1.40***	−0.72	−1.07	−0.66	**−2.32***	**−1.78***	**−2.74***
BiAraLI-03	5;7	raw score	25	4	35	24	10	13	24
		z-score	**−3.28***	**−2.40***	−1.13	−0.58	−0.67	0.36	−0.35
BiAraLI-04	6;0	raw score	47	28	47	28	12	14	24
		z-score	−0.20	−0.49	0.01	−0.31	−0.54	0.72	−0.68
BiAraLI-05	7;3	raw score	43	23	46	35	13	14	27
		z-score	**−2.60***	**−1.53***	−0.86	−0.40	−0.59	0.61	0.02
BiAraLI-06	5;4	raw score	33	12	49	33	5	3	13
		z-score	**−2.07***	**−1.73***	0.41	0.34	**−2.06***	**−4.67***	**−3.26***
BiAraLI-07	6;1	raw score	46	26	39	26	15	11	25
		z-score	−0.34	−0.64	−0.78	−0.49	0.22	−0.53	−0.39
BiAraLI-08	7;1	raw score	55	35	36	10	9	11	22
		z-score	0.71	−0.23	**−2.03***	**−2.62***	**−1.76***	−1.08	**−1.92***
BiAraLI-09	6;7	raw score	48	31	57	39	18	12	29
		z-score	−0.07	−0.26	0.99	0.64	0.98	−0.11	0.79
BiAraLI-10	5;0	raw score	21	0	35	23	10	7	13
		z-score	**−3.89***	**−2.73***	−1.13	−0.69	−0.67	**−2.66***	**−3.26***
BiAraLI-11	6;4	raw score	51	41	47	24	6	11	15
		z-score	0.33	0.49	0.01	−0.66	**−2.07***	−0.53	**−3.33***

Note. Z-scores below −1.25 are in boldface and marked with *.

In Figure 3, age-adjusted z-scores are plotted for Arabic and Swedish vocabulary comprehension (a) and production (b) for the children in the TD and DLD samples. As is evident in Table 11 and Figure 3, more than half (7/11) of the children in the DLD group had a z-score below −1.25 for comprehension in one language, but only one child (BiAraLI-01) received a z-score below −1.25 for comprehension in both languages. The pattern was similar for vocabulary production scores; six of the children in the DLD group scored below −1.25 in one language (additionally, one child, BiAraLI-01, scored at the cut-off in Arabic). Thus, it was not the case that all or even the majority of the DLD children scored below −1.25 in either task or in both languages. For both comprehension and production, there was a notable overlap in scores for the children in the DLD sample and children in the TD sample.

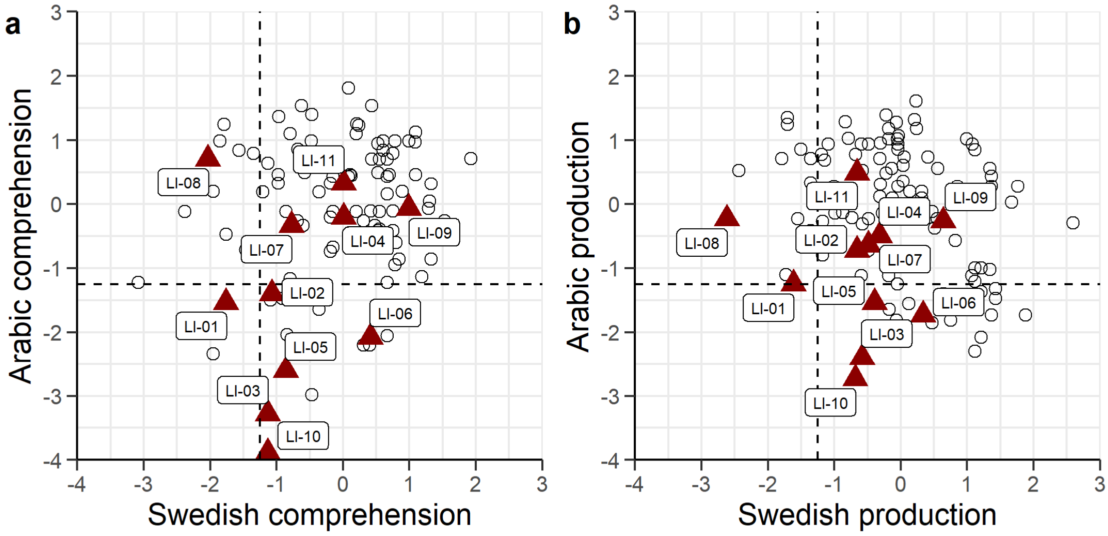

Figure 3. Scatterplots showing age-adjusted z-scores of (**a**) Arabic and Swedish vocabulary comprehension and (**b**) Arabic and Swedish vocabulary production of the children in the DLD sample (triangles and labels) compared to the children in the TD sample (circles). Dashed lines at −1.25.

In Figure 4, z-scores are plotted for the three NWR tasks for the children in the TD and the DLD samples, for the (a) LS-Swe task, (b) CL-Swe task, and (c) NWRT-Leb. As can be seen in Table 11 and Figure 4, there was a notable overlap in NWR performance in the two samples.

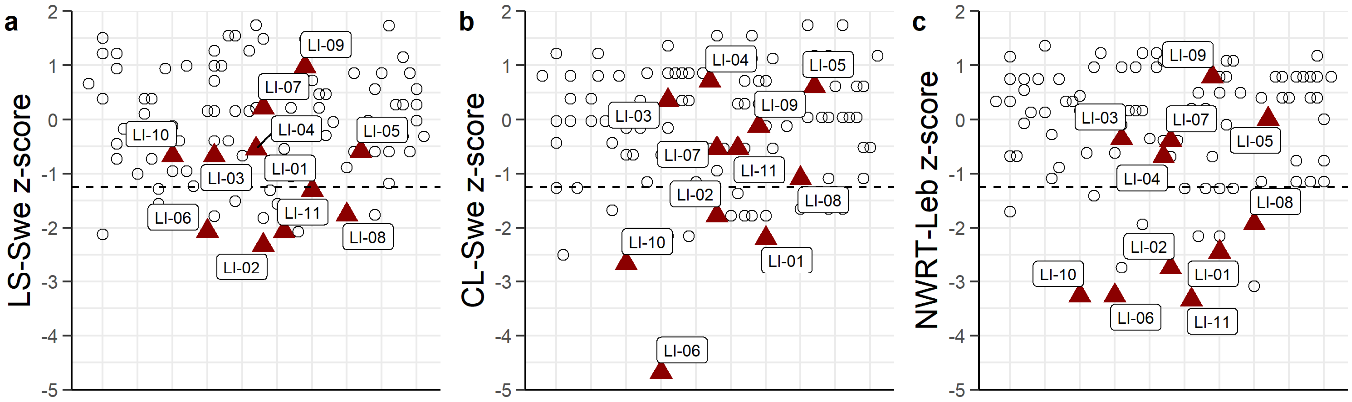

Figure 4. Scatterplots showing age-adjusted z-scores of the (**a**) LS-Swe, (**b**) CL-Swe and (**c**) NWRT-Leb tasks of the children in the DLD sample (triangles and labels) compared to the children in the TD sample (circles). Dashed lines at −1.25.

We have refrained from calculating sensitivity and specificity for the vocabulary and the NWR tasks since, as the scatterplots in Figures 3 and 4 show, there is a lot of overlap between the TD and the DLD groups and we see no straightforward solution for getting around this overlap (e.g., by exploring different cut-offs).

To summarise, although the children in the DLD sample typically scored below the mean on both vocabulary comprehension and production in both languages, it was rare that they scored low (i.e., below −1.25 z-scores) in *both* languages. For non-word repetition, performance was also generally below the mean score of their age peers, but only six DLD children scored below −1.25 (in z-scores) in at least one task. We will now investigate whether the DLD children's performance on certain types of NWR items differs more from their TD peers.

3.3.2. Performance of the TD and the DLD Children on NWR Items with Different Properties

In this section, NWR accuracy (% correct responses) in the TD sample and the DLD sample will be compared by item length (syllables) and presence of clusters separately for each task. Since the age range was narrower in the DLD sample (5;0–7;3) compared to the TD sample (4;0–7;11), we excluded the four-year-old TD children here. Due to the large difference in sample size between the two groups (N_{TD} = 77, N_{DLD} = 11), only descriptive statistics will be reported.

As evident in Table 12, at group level, the DLD children scored lower than the TD children on all tasks, at all item lengths, and for items with and without clusters alike, with two exceptions. For the shortest LS-Swe items (two syllables) with clusters, accuracy was higher for the DLD group (91%) compared to the TD group (84%), and for the shortest NWRT-Leb items (one syllable) without clusters, accuracy was at ceiling (TD = 99%, DLD = 100%) for both groups. We could not discern any tendencies for the DLD children to perform *disproportionally worse* on one type of NWR task or on certain types of items, such as long items and/or phonologically complex items (with clusters).

Table 12. Accuracy (% correct answers) for NWR tasks by number of syllables and presence of clusters for the 5–7-year-olds in the TD sample (N = 77) and the children in the DLD sample (N = 11).

	LS-Swe				**CL-Swe**		**NWRT-Leb**			
	w/o Clusters		*With Clusters*				*w/o Clusters*		*With Clusters*	
	TD	*DLD*	*TD*	*DLD*	*TD*	*DLD*	*TD*	*DLD*	*TD*	*DLD*
1 syllable	–	–	–	–	–	–	99%	100%	98%	94%
2 syllables	80%	75%	84%	91%	98%	91%	95%	79%	86%	70%
3 syllables	82%	52%	46%	30%	92%	77%	80%	49%	74%	38%
4 syllables	84%	64%	26%	6%	82%	59%	–	–	–	–
5 syllables	46%	27%	16%	0%	40%	25%	–	–	–	–

In the next section, the DLD children's test results will be analysed in conjunction with background information and reports about functional language abilities from parents, teachers, and SLPs.

3.3.3. Background Information and Functional Language Abilities of the Children in the DLD Sample

According to parental report, 9/11 of the children had a delayed language development in their first language (Arabic), and 6/11 had a late development in their second language (Swedish). Six children in the DLD sample (55%) produced their first word and/or first word combination late, i.e., later than 12 and 24 months, respectively, and the same proportion of children (55%) had a close relative with language or literacy problems. By comparison, 27% of the children in the TD sample were reported to have a late onset of the first word or word combination, and only 7% had a close relative with language or literacy difficulties.

The reports from parents, teachers and SLPs regarding functional language skills and communicative behaviour of the children in the DLD sample are briefly summarised in Table A2 in the Appendix A. Most children in the DLD sample had difficulties with both comprehension (e.g., having difficulties understanding instructions) and production (e.g., being difficult to understand and making oneself understood, having deficits in expressive morphosyntax or speaking in rudimentary utterances). The children in the DLD sample will now be characterised in terms of their performance on the vocabulary and NWR tasks and discussed in light of the information provided by parents, teachers, and SLPs.

BiAraLI-03 and BiAraLI-05 had poor vocabulary scores (particularly in Arabic), but received positive or only slightly negative z-scores in the NWR tasks. At the same time, both children had functional communication difficulties according to the parents, SLPs, and teachers (although the mother of BiAraLI-05 thought that his Arabic was age-appropriate). Two children (BiAraLI-04 and BiAraLI-07) scored moderately low in both vocabulary and NWR tasks. In comparison to the other children in the DLD group, these two children had seemingly milder problems. BiAraLI-04 recently had his diagnosis changed from general to expressive language disorder. Although the SLP reported deficits in expressive morphosyntax in both languages, the parents and the preschool staff were of the opinion that he only had problems with pronunciation, and the preschool staff thought that he had a small Swedish vocabulary due to poor exposure. BiAraLI-07 had problems with both comprehension and production according to the parents, the SLP, and the preschool teacher. However, it was reported that he played well with other children with few misunderstandings or conflicts.

Five children (BiAraLI-01, BiAraLI-02, BiAraLI-06, BiAraLI-08, and BiAraLI-10) performed poorly on all NWR tasks and had poor vocabulary scores in one or both languages. Four of these children were described by parents, SLPs, and teachers alike as having *severe* language difficulties, to the extent that there were frequent misunderstandings or conflicts with peers. Reports about the fifth child (BiAraLI-06) were mixed, as the parents and the preschool teacher did not find his language skills to be severely affected, but the SLP reported great difficulties in several language domains. BiAraLI-11 had a large discrepancy between performance in the vocabulary tasks and the NWR tasks. She had good vocabulary scores in Arabic, and surprisingly good scores in Swedish considering that her age of onset for Swedish was late (age 4;0–5;0). However, NWR performance was poor, especially in the tasks with higher phonological complexity (LS-Swe and NWRT-Leb). Her comprehension seemed to be better than production according to parent, SLP and teacher interviews.

Finally, BiAraLI-09 scored high or very high in all NWR tasks. His vocabulary scores were slightly below the mean in Arabic and far above the mean in Swedish. The reports from parents, the SLP, and the school were inconsistent. The parents said that they were concerned about his early language development, and the child had been seen by different SLPs during the course of a couple of years. Eventually, he was diagnosed with DLD at the SLP clinic, but the parents were not sure that the diagnosis was accurate. At the same time, the school staff perceived the child to be very shy and reported that he did not like to speak in class, but that he seemed to have age-appropriate expressive skills during individual sessions with the special education teacher. Considering that BiAraLI-09 had high scores in all NWR tasks, very high vocabulary scores in Swedish, and Arabic scores just below the TD mean, it could be the case that this child was subject to overdiagnosis of DLD. According to the SLP, he had only been assessed in Swedish and his language scores were compared against monolingual Swedish norms.

4. Discussion

In this study, we investigated vocabulary comprehension and production, as well as the NWR performance of 110 Arabic-Swedish-speaking bilinguals aged 4–7. The relative effect of age, language exposure and SES on vocabulary comprehension and production was investigated for the minority and the majority language. NWR performance was investigated in relation to age, language exposure, vocabulary, and properties of the nonword items. We also explored whether bilingual children with a diagnosis of DLD could be distinguished from TD children, based on their performance on vocabulary and NWR tasks, and whether one particular type of NWR task might identify DLD better.

4.1. Vocabulary in the TD Sample

We found that vocabulary comprehension and production scores increased with age in both the minority language Arabic and the majority language Swedish. These results accord with findings reported in the literature, namely that there is a clear development with age in the majority language (Bialystok et al. 2010; Cobo-Lewis et al. 2002a; Prevoo et al. 2014). However, our finding of a clear development with age in Arabic differs from many previous studies that report small or no gains with age in the minority language (Ganuza and Hedman 2019; Gathercole and Thomas 2009; Leseman 2000; Öztekin 2019, chp. 4). Recall that a bit less than half of the children (42%) were not born in Sweden, 48% had an age of onset to Swedish after age 3, and 20% had less than two years of exposure to Swedish at the time of testing. As mentioned in the introduction, Arabic speakers are the largest linguistic minority in Sweden, and many children in our sample had several sources of input outside the home (e.g., in (pre)school or at community centres arranging activities for Arabic-speaking children). This means that at group level, the children in our study had a high amount of cumulative exposure in Arabic from various interlocutors, which probably supported their development of the minority language.

Vocabulary scores increased as a function of the proportion of daily exposure. The effect was seen for both languages, and it was stronger for production than for comprehension. These findings are in line with earlier studies demonstrating a relationship between the relative amount of exposure and vocabulary comprehension in the minority language (Prevoo et al. 2014) and the majority language (Unsworth 2016), as well as for vocabulary production in the majority language (Öztekin 2019, chp. 4; Prevoo et al. 2014). They are also in line with Thordardottir's (2011) observation that the effect of relative amount of exposure is stronger for vocabulary production than comprehension. In the present study, language exposure was further investigated by exploring the effect of length of exposure (LoE) (in months). For the minority language Arabic, LoE could not be investigated separately as it coincided with chronological age. For the majority language Swedish, LoE emerged as the most influential predictor of vocabulary scores, overshadowing age and daily exposure. Interestingly, these results go against those of Thordardottir (2019), who found that LoE was not a significant predictor of vocabulary comprehension scores in slightly older children (Canadian 7–9-year-olds with French as their common language) when cumulative exposure was also accounted for. Since the present study measured current amount of exposure as a separate variable, our length of exposure variable is likely to capture length as well as cumulative (amount of) exposure.

There was no effect of SES (parental education) on vocabulary comprehension or production, neither in the minority language Arabic nor in the majority language Swedish. Considering previous reports of null results for SES and vocabulary in the minority language (Cobo-Lewis et al. 2002b; Öztekin 2019, chp. 4; Prevoo et al. 2014), it was unsurprising that SES was not a significant predictor of Arabic vocabulary. Surprisingly however, SES was not a significant predictor of Swedish vocabulary either. This result does not match previous studies from other countries, where higher SES is generally associated with better vocabulary scores in the majority language (Buac et al. 2014; Calvo and Bialystok 2014; Cobo-Lewis et al. 2002a; Prevoo et al. 2014). There may be several reasons for why SES was not a significant predictor of majority language vocabulary. Higher SES tends to co-vary with a higher degree of majority language use in the home (Prevoo et al. 2014), or better majority language proficiency of the parents (Buac et al. 2014), which may in turn boost the child's majority language skills if the parents speak the majority language in the home. In the present study, almost 80% of the participating families reported that both parents spoke to their child only or mainly in the minority language Arabic. Overall, there was very little parental input in Swedish, and it is therefore unlikely to boost majority language vocabulary in our sample of Arabic-Swedish-speaking bilinguals. Yet another explanation could be the Swedish setting, where access to institutional childcare (preschool) is not dependent on family income or SES (as it often is in other countries). Consequently, some differences between children from different SES backgrounds may be levelled out, and their vocabulary development may be influenced more strongly by the quantity and quality of input in (pre)school and language-fostering activities in the home, which in turn are not directly related to the parents' educational level (Bohnacker et al. 2021).

4.2. Non-Word Repetition in the TD Sample

In the present study, we found age effects for all three types of NWR tasks: the LS-Swe, the CL-Swe, and the NWRT-Leb. Scores increased with age, mirroring findings from several previous studies (Chiat and Roy 2007; Kalnak et al. 2014; Radeborg et al. 2006). These age effects held for all tasks also when controlling for vocabulary scores and for length of exposure to Swedish (for the LS-Swe task) in the multivariate regression models. For all three NWR tasks, accuracy decreased with increased non-word length, in line with earlier studies using other stimulus items (Boerma et al. 2015; Chiat and Roy 2007; Dollaghan and Campbell 1998; Ellis Weismer et al. 2000; Thordardottir and Brandeker 2013). Increased phonological complexity (presence of consonant clusters) had an adverse effect on repetition accuracy in the two tasks that contained items with clusters (LS-Swe and NWRT-Leb). This is in line with previous studies reporting lower accuracy for phonologically more complex items (Abed Ibrahim and Hamann 2017; dos Santos and Ferré 2018; Jones et al. 2010). It was not the case though that items with higher phonological complexity were generally more difficult to repeat. Task difficulty differed greatly; the two tasks that contained items with clusters had the overall highest repetition accuracy (NWRT-Leb) and the lowest (LS-Swe). Interestingly, accuracy for length (number of syllables) and presence of clusters was very different for all NWR tasks. For instance, accuracy of the LS-Swe items decreased by similar amounts for each added syllable. By contrast, for the CL-Swe task, accuracy decreased only slightly for each added syllable between 2–4 syllables, but then declined abruptly at five syllables. Accuracy on the NWRT-Leb was only marginally lower for items with vs. without clusters at all syllable lengths, whilst for the LS-Swe task, there were large discrepancies in accuracy for items with vs. without clusters at different syllable lengths. For 2-syllable items with clusters, accuracy was even somewhat higher than for items without clusters, but much lower for items with 3–5 syllables. Recall however that all tasks utilized different phoneme inventories, the LS-Swe having a wide variety of language-specific Swedish phonemes, whilst the NWRT-Leb had a very restricted phoneme inventory. We speculate that there is an interplay between phoneme inventory, syllabic complexity and item length (number of syllables), affecting item and overall task difficulty.

There was no correlation between any of the tasks and SES, mirroring several previous studies that did not find an association between SES and NWR performance (Boerma et al. 2015; Chiat and Roy 2007; Kalnak et al. 2014). Daily exposure did not correlate with performance on the LS-Swe task (daily exposure to Swedish), the CL-Swe (daily exposure to Swedish) nor the NWRT-Leb (daily exposure to Arabic). Furthermore, length of exposure to Swedish was not a significant predictor of performance on the LS-Swe task when chronological age was controlled for. Swedish vocabulary (comprehension) was a significant predictor of performance for the LS-Swe items. This is congruent with several previous studies finding an association between NWR performance and vocabulary (Gibson et al. 2015; Kohnert et al. 2006; Sorenson Duncan and Paradis 2016; Thordardottir and Brandeker 2013). Interestingly, there was a vocabulary effect on all three NWR tasks. Contrary to expectation, the effect of Swedish vocabulary on repetition accuracy was not stronger for the language-specific LS-Swe items than for the non-language-specific CL-Swe items. Additionally, the impact of Arabic vocabulary on NWRT-Leb is somewhat surprising as this task was constructed to be language-independent and to minimize the impact of vocabulary (dos Santos and Ferré 2018).

Finally, as many as 30% of the children in the TD sample (4-, 5- and 6-year-olds) scored below −1 SD on the language-specific LS-Swe task compared to monolingual reference data (Radeborg et al. 2006). Interestingly, this proportion is similar to that reported by Sorenson Duncan and Paradis (2016), who found that 29% of the bilingual Canadian children in their sample scored below −1 SD on a language-specific English NWR task and, unlike the present study, they found that performance was affected by (cumulative) exposure to English. Thus, even though NWR performance on the LS-Swe task was not measurably related to language exposure *per se*, it cannot be assumed that the NWR performance of bilingual children can be compared against monolingual norms.

4.3. Vocabulary and Non-Word Repetition in the DLD Sample

Vocabulary was assessed in both Arabic and Swedish. At group level, the children in the DLD group scored below the cross-sectional mean, but within the range for their age group on comprehension and production in Arabic and Swedish. There were, however, large individual differences, and not all children had poor scores in both languages. While only one child had a z-score below −1.25 in both languages (in both modalities in Swedish but only in comprehension in Arabic), five children had a z-score below −1.25 in Arabic comprehension, and four children in Arabic production. Thus, having a z-score below −1.25 in *both* languages may not be a valid criterion for identifying DLD in this group of children. As described in the introduction, it is frequently argued that language difficulties must show in both languages of bilingual children (Kohnert 2010; Salameh et al. 2002; Thordardottir 2015), but evidence-based recommendations for interpreting language test scores and choosing suitable cut-offs are rare. Peña et al. (2016) investigated whether cut-offs established for monolingual populations could accurately classify Spanish-English bilinguals with balanced exposure to both languages as DLD or TD on a task targeting semantic skills. They found that scoring below the monolingual cut-off in both languages correctly classified the children as DLD, whereas considering only one language led to overidentification. In the current study, there was much variation in age of onset to Swedish and in the proportion of relative exposure to each language. Notably, most DLD children with poor vocabulary scores had low scores in Arabic, despite the fact that they had received continuous exposure to the language from birth. In light of this, we propose that having low vocabulary scores in the home language *despite* early onset and continuous input may be a warning sign for DLD.

Let us now move onto NWR, as this has been described as a promising diagnostic tool for bilingual children. At group level, the children in the DLD group scored below the mean of their TD age peers on most NWR tasks. However, there was much individual variation and only six (out of 11) DLD children had a z-score below −1.25 in at least one NWR task. Additionally, there was considerable overlap between the TD and the DLD groups on all tasks, with some DLD children scoring above the mean and some TD children scoring below the −1.25 z-score cut-off. Although poor NWR performance is frequently reported to be associated with DLD (Boerma et al. 2015; Dollaghan and Campbell 1998; Kalnak et al. 2014), there is not a perfect relationship between low NWR scores and presence of DLD (Ellis Weismer et al. 2000). Moreover, there are several reports in the literature of poorer diagnostic accuracy and a higher degree of overlap between TD and DLD groups on NWR tasks in bilingual populations compared to monolingual populations. Poorer diagnostic accuracy in bilinguals has not only been attested for language-specific NWR tasks, but also for language-non-specific and quasi-universal tasks (Abed Ibrahim and Hamann 2017; Boerma et al. 2015; dos Santos and Ferré 2018; Schwob et al. 2021; Thordardottir and Brandeker 2013). In the current study, the DLD children did not perform disproportionally worse on a certain task or item type compared to the TD children. Rather, the DLD children generally scored lower than the TD children on all item types, and accuracy decreased for both groups with increased item length and the presence of consonant clusters. These findings are in line with previous studies reporting that both TD and DLD children have more difficulty with NWR as stimulus length increases (Boerma et al. 2015; Schwob et al. 2021). At the same time, results are mixed with regard to whether DLD children have disproportionally more difficulties with longer items and/or phonologically complex items compared to TD children (Boerma et al. 2015; Schwob et al. 2021). In conclusion, we did not find that one type of NWR task or items with certain properties were superior in the identification of DLD.

Two conclusions can be drawn from these observations about vocabulary and NWR performance in our sample of Arabic-Swedish bilinguals. The first is that bilinguals with DLD do not necessarily perform low in both languages, even when comparing them to peers who grew up in the same country, speaking the same language combination. The second is that performance on NWR tasks cannot reliably distinguish all children with a DLD diagnosis when comparing them to a large group of children with (according to parental report) typical language development. Bearing in mind that vocabulary is the linguistic domain that is probably the most affected by language exposure, perhaps it is not surprising to find a large overlap in vocabulary scores, as fluctuating patterns of exposure give rise to much variation in performance on vocabulary tasks in both the TD and the DLD groups. However, our study indicates that claims such as "bilinguals with DLD must perform low in both languages in order to qualify for a DLD diagnosis" must be taken with caution. In the area of vocabulary and non-word repetition, they are clearly not supported by the empirical evidence.

It is noteworthy that NWR performance was clearly poor in only half of the children with a DLD diagnosis, with much overlap in performance between the DLD and TD groups. As Norbury et al. (2016) point out, cut-offs are arbitrary in the sense that they do not say anything about how a certain score corresponds to functional communicative abilities. Thus, receiving a z-score below −1.25 on a given task is not necessarily associated with poor functional language skills. Conversely, individuals who score above (and in some cases well above) −1.25 on a certain task do not necessarily have sufficient functional language skills. Notably, there was subgroup of around half of the children in the DLD sample who were described by both parents, teachers and SLPs as having *severe* communication difficulties that often led to peer conflicts and had a negative impact on their learning outcomes. These children scored low on the majority of the NWR tasks. However, there was also another subgroup in the DLD sample who were described to have somewhat milder language problems, and who scored low but still within the typical range on the NWR tasks. In a clinical setting, for the most severe cases (like the children in the first subgroup), a language disorder is usually not difficult to determine. The difficult cases are rather those falling into the second category. NWR does not seem to have good diagnostic accuracy for the bilingual children in our study; it is at best suggestive. Therefore, we argue that it is crucial to *interpret* language test scores in light of a detailed case history and reports from parents and teachers.

As is well known, the initial diagnosis of DLD is generally difficult in bilinguals, with a risk of misdiagnosis by the experts. The children with a DLD diagnosis in our sample had undergone careful and often repeated assessment, sometimes by several SLPs. However, we cannot completely rule out that there may have been some misdiagnosis, particularly in one case (BiAraLI-09). Furthermore, as DLD is a heterogeneous condition, individual children may have relative strengths or weaknesses in one modality (comprehension or production) or one or more linguistic domains (phonology, morphosyntax, vocabulary, discourse or pragmatics), which may explain why some of the DLD children scored unexpectedly high on a certain task.

In the present study, a much higher proportion of children in the DLD sample had a late language development and/or heredity for speech, language or literacy problems. These findings accord with earlier research showing that delayed language development and heredity for disorders of language or literacy are disproportionally more common in children with DLD compared to their TD peers (Kalnak et al. 2012; Paradis et al. 2010; Trauner et al. 2000). Parents, teachers and SLPs were asked to characterise the children's language and communication. Most children in the DLD sample were described as having deficits in their functional language skills. Descriptions of poor language comprehension were common (e.g., having difficulties understanding instructions or complex syntax), which is frequently reported in the literature about children with DLD (Bishop 1997; Friedmann and Novogrodsky 2004; Norbury et al. 2016). The children were also reported to have poor expressive abilities, for instance having deficits in expressive morphosyntax, which is also a common feature among children with DLD (Paradis et al. 2022; Reuterskiöld et al. 2021). In conclusion, we found that it was necessary to combine a formal assessment of vocabulary and NWR with a detailed case history and reports from parents, teachers and SLPs about early language development, heredity for language and literacy problems, language exposure, and functional language skills for identifying language difficulties, particularly in those children who performed borderline poor on NWR. This solution is also supported by other studies finding that combining NWR with parental questionnaires probing early language milestones and parental concern about the child's language development can improve diagnostic accuracy (Boerma and Blom 2017; Paradis et al. 2013).

5. Conclusions

For this understudied language combination of Arabic-Swedish-speaking bilinguals, we found that language exposure had a large impact on minority and majority vocabulary scores, but SES was not a significant predictor of vocabulary in any language. Age and vocabulary size had a positive effect on NWR performance; longer items and items containing clusters had lower repetition accuracy. A language-specific Swedish NWR task was evaluated for the first time for a large group of bilinguals, and results showed that although language exposure did not measurably affect NWR scores, these bilinguals were disadvantaged when compared against monolingual norms. There was a substantial overlap between TD and DLD children in performance on both vocabulary and NWR tasks. Low vocabulary scores in the minority language *despite* ample and continuous exposure from birth emerged as a warning sign for DLD. Diagnostic accuracy seemed at best suggestive for NWR, and we could not discern any particular task or type of item that was clearly superior for identifying DLD in our sample. Most children with DLD did not score below the −1.25 z-score cut-off in both languages (for vocabulary), and many scored above this cut-off on a majority of the NWR tasks. Reports from parents and teachers on language exposure, language development, concerns, functional language skills and communication difficulties are crucial when assessing suspected DLD in bilinguals. Future research should include a larger group of DLD children, as well as use longitudinal designs in order to investigate and confirm the results we have shown here for our cross-sectional sample.

Author Contributions: Conceptualization: U.B. and L.Ö.; Data curation: U.B. and L.Ö.; Formal Analysis: U.B. and L.Ö.; Funding Acquisition: U.B.; Investigation: U.B. and L.Ö.; Methodology, U.B. and L.Ö.; Project administration: U.B.; Resources: U.B.; Supervision: U.B.; Validation: U.B.; Visualisation: L.Ö.; Writing—original draft preparation: U.B. and L.Ö.; Writing—review and editing: U.B. and L.Ö. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Swedish Research Council [VR 431-2013-1309] and the Bank of Sweden Tercentenary Foundation [P19-0644:1].

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, Swedish regulations concerning data protection and research ethics, and the local university ethical code of conduct. The Faculty of Languages at Uppsala University does not implement an institutional review and approval process.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Participation could be discontinued at any time.

Data Availability Statement: The present study is part of a larger, ongoing research project. For data supporting the reported results, please consult the authors.

Acknowledgments: The authors extend their thanks to Rima Haddad who collected, transcribed and scored the Arabic CLT and NWR data.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. All items in the NWR tasks: the LS-Swe (language-specific Swedish), the CL-Swe (crosslinguistic Swedish), and the NWRT-Leb (Non-word repetition task Lebanese).

LS-Swe	/ɡlʏˈvoː/ /aˈpɛt/ /ɪˈfʉːm/ /ˈɧɔrjɛ/ /naˈkiːt/ /ˈspʉːmɛ/ /lɛbʊˈsʉːf/ /mɵstrɛˈfalj/ /ɡlɛŋɛˈsɵlp/ /salʊˈtɑːn/ /hœntˈpʉːlɛ/ /nɛsʊˈloː/ /ˈmaŋɛʂˌblɛɡɛ/ /ɛlʊˈmɔkɪ/ /ɔlɪˈtʉːkɛ/ /spɵrɪfraˈɡoːl/ /tɪbɛˈfiːmɛ/ /lɵtʊspɛˈlʉːn/ /tœlɪpaˈleːrʊ/ /ɕɵlɛˌkrɔmpaˈmiːd/ /fɪmɪɡlaˈnɛftɪ/ /hɪlʊteraˈpʉːd/ /flɛtɛˌmɪŋɛˈroːf/ /dalabɛlˈhiːmɛ/
CL-Swe	/sɪbʊ/ /dʊla/ /nagɪ/ /lʊnɪ/ /sɪpʊla/ /banʊdɪ/ /malɪtʊ/ /lɪmɪka/ /sɪbalɪta/ /mʊkɪdala/ /ɡasʊlʊmɪ/ /lɪdɪsakʊ/ /sɪpʊnakɪla/ /tʊlɪɡasʊmʊ/ /malʊsɪɡʊba/ /lɪdapɪmʊtɪ/
NWRT-Leb	/fla/ /lafi/ /kafib/ /flablu/ /bufaki/ /fuk/ /kib/ /bukli/ /kifabu/ /blaklu/ /kuflabi/ /baf/ /faku/ /kabufik/ /flukif/ /blufa/ /kubafli/ /biklafu/ /fablu/ /bukif/ /fliku/ /klu/ /fikubla/ /bli/ /klifak/ /bilu/ /flibuka/ /kabi/ /bifakub/ /klibafu/

Table A2. Language difficulties and behaviour of the DLD children as described by parents, teachers, and SLPs.

		Parents	**Teacher**	**SLP**
BiAra LI-01	*language*	– comp, prod	– comp, prod, vocab	– comp, prod, vocab
	behaviour/ comments	LLD, rarely interacts with other children	frequent misunderstandings	minimal progress in therapy
BiAra LI-02	*language*	– comp, prod	– comp, prod	– comp, prod, vocab
	behaviour /comments	LLD	peer conflicts, easily distracted	responds well to therapy, limited attention span
BiAra LI-03	*language*	– comp, prod	– comp, prod, vocab	– comp, prod
	behaviour/ comments	LLD, sometimes gets angry if not understood	unfocused and has trouble listening	rarely initiates communication
BiAra LI-04	*language*	– pronunciation + comp, prod	– vocab, pronunciation + comp, prod	– vocab, prod, expr. phonology ok comp
	behaviour/ comments	talkative, great social skills	plays well with other children	has made great progress in therapy
BiAra LI-05	*language*	+ comp, prod	– comp, prod	– comp, prod, vocab
	behaviour /comments	LLD, no problems in Ara, very little Swe exposure	rarely speaks, uses gestures to make himself understood	attention difficulties during assessment
BiAra LI-06	*language*	– prod + comp	– vocab ok comp, prod	– comp, prod, pragmatics
	behaviour/ comments	LLD, now learns fast	meticulous about rules and routines	talks a lot, but has difficulties getting his meaning across
BiAra LI-07	*language*	– vocab, word-finding, pronunciation	– comp, prod, vocab	– comp, prod, vocab
	behaviour /comments	LLD, now learns fast	plays well with other children	responds well to therapy, some attention difficulties
BiAra LI-08	*language*	– comp, prod	– comp, prod	– comp, prod, vocab
	behaviour/ comments	only interacts with Ara-speaking children	peer conflicts, only interacts with Ara-speaking children	SLP recommends psychological assessment
BiAra LI-09	*language*	– vocab (Swe) + comp, prod (Ara)	+ comp + prod (in small groups)	– comp, prod, vocab
	behaviour /comments	LLD, shy, seldom initiates contact with peers	shy, does not like to speak in class	shy and cautious during assessment

Table A2. *Cont.*

		Parents	Teacher	SLP
BiAra LI-10	*language*	– comp, prod	– comp, prod	– comp, prod, vocab
	behaviour/ comments	LLD, likes to interact with other children	frequent misunderstandings, teachers must often act as 'interpreters'	talks a lot, but is very difficult to understand (incomplete utterances)
BiAra LI-11	*language*	– prod, pronunciation + comp	– comp, prod	– comp, prod, pronunciation
	behaviour/ comments	LLD, self-conscious and avoids speaking	unwilling to speak, does not initiate play	shy, self- conscious and avoids speaking

Note. LLD = Late language development; 'comp' = language comprehension; 'prod' = language production; '–' indicates relative weaknesses, '+' indicates relative strengths in language ability.

Notes

[1] Other terms such as (Specific) Language Impairment ((S)LI) and, less commonly, Primary Language Impairment (PLI) are used in the literature to refer to the same condition. DLD will be used throughout for consistency.

[2] Note that the 'crosslinguistic' NWR task was previously called 'the quasi-universal nonword repetition test' (https://www.bi-sli.org/cl-nonword-repetition); (e.g., Chiat 2015; Boerma et al. 2015).

[3] Note that this task has recently been renamed 'the quasi-universal nonword repetition test' (https://www.bi-sli.org/qu-nonword-repetition) but was previously referred to as 'the LITMUS-NWR' (e.g., de Almeida et al. 2017; dos Santos and Ferré 2018; Tuller et al. 2018) or 'the LITMUS-NWRT' (e.g., Abed Ibrahim and Fekete 2019; Abed Ibrahim and Hamann 2017).

[4] This also includes three single-parent households where data was available for that single parent only.

[5] Note that there is no standardised procedure or clinical guidelines for assessing suspected language disorders in bilinguals in Sweden.

[6] For a detailed description of the adaptation procedure, see Bohnacker et al. (2021).

[7] Since the CL-Ara results were generally similar to the CL-Swe task, they are not reported here. For results on the CL-Ara task, see Öberg (2020).

[8] For one seven-year-old, data was missing, and for one five-year-old, AoO to Arabic was at age 1.

[9] Descriptive statistics are reported here for age groups so that they can be used as reference data. For comparisons between the two languages (Arabic and Swedish), modalities (comprehension and production), and age groups, see Öberg (2020) and Bohnacker et al. (2021).

[10] Descriptive statistics are reported here for age groups so that they can be used as reference data. Age-group comparisons are available in Öberg (2020).

[11] The low performance on the LS-Swe task is also noteworthy from another perspective. This is the only NWR task for which published reference data exists (for monolingual Swedish children age 4–6, N = 200, Radeborg et al. 2006). Compared to the monolingual reference data, the proportion of the bilinguals in the present study that scored -1 SD from the mean was 30% (22/73). This is twice as many as expected if performance were the same for monolinguals and bilinguals.

[12] We interpret this finding as follows: in this maximal model, chronological age and properties of the items overshadow the (relatively smaller) effect of vocabulary. Note that in Model 7, Swedish vocabulary is a predictor of NWR accuracy for both LS-Swe and CL-Swe items.

References

Abed Ibrahim, Lina, and Cornelia Hamann. 2017. Bilingual Arabic-German and Turkish-German Children with and without Specific Language Impairment: Comparing Performance in Sentence and Nonword Repetition Tasks. In *Proceedings of the 41st Annual Boston University Conference on Language Development*. Edited by Maria LaMendola and Jennifer Scott. Somerville: Cascadilla Press, pp. 1–17.

Abed Ibrahim, Lina, and István Fekete. 2019. What Machine Learning Can Tell Us about the Role of Language Dominance in the Diagnostic Accuracy of German LITMUS Non-Word and Sentence Repetition Tasks. *Frontiers in Psychology* 9: 2757. [CrossRef] [PubMed]

Abou Melhem, Nouhad, Edith Kouba Hreich, and Christophe dos Santos. 2011. *The Non-Word Repetition Task-Lebanese (Adapted from the NWR-FRENCH Task, Now Part of the LITMUS Battery)*, Unpublished material.

Andersson, Ketty, Kristina Hansson, Ida Rosqvist, Viveka Lyberg Åhlander, Birgitta Sahlén, and Olof Sandgren. 2019. The Contribution of Bilingualism, Parental Education, and School Characteristics to Performance on the Clinical Evaluation of Language Fundamentals: Fourth Edition, Swedish. *Frontiers in Psychology* 10: 1586. [CrossRef] [PubMed]

Archibald, Lisa M. D. 2008. The Promise of Nonword Repetition as a Clinical Tool. *Canadian Journal of Speech-Language Pathology and Audiology* 32: 21–28.

ASHA. 2004. Preferred Practice Patterns for the Profession of Speech-Language Pathology. *Preferred Practice Patterns*. [CrossRef]

Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Barthelom, Eleonora, and Monika Åkesson. 1995. Konstruktion, Testning och Utvärdering av Nonord. Master's thesis, Lunds Universitet, Lund, Sweden.

Bartoń, Kamil. 2020. MuMIn: Multi-Model Inference. R Package Version 1.43.17. Available online: https://CRAN.R-project.org/package=MuMIn (accessed on 15 February 2022).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]

Bialystok, Ellen, Gigi Luk, Kathleen F. Peets, and Sujin Yang. 2010. Receptive Vocabulary Differences in Monolingual and Bilingual Children. *Bilingualism: Language and Cognition* 13: 525–31. [CrossRef]

Bishop, Dorothy V. M. 1997. *Uncommon Understanding. Development and Disorders of Language Comprehension in Children*. Hove: Psychology Press.

Boerma, Tessel, and Elma Blom. 2017. Assessment of Bilingual Children: What If Testing Both Languages Is Not Possible? *Journal of Communication Disorders* 66: 65–76. [CrossRef]

Boerma, Tessel, Frank Wijnen, Paul Leseman, and Elma Blom. 2017a. Grammatical Morphology in Monolingual and Bilingual Children with and without Language Impairment: The Case of Dutch Plurals and Past Participles. *Journal of Speech, Language and Hearing Research* 60: 2064–80. [CrossRef] [PubMed]

Boerma, Tessel, Paul Leseman, Frank Wijnen, and Elma Blom. 2017b. Language Proficiency and Sustained Attention in Monolingual and Bilingual Children with and without Language Impairment. *Frontiers in Psychology* 8: 1241. [CrossRef]

Boerma, Tessel, Shula Chiat, Paul Leseman, Mona Timmermeister, Frank Wijnen, and Elma Blom. 2015. A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children Learning Dutch as a Second Language. *Journal of Speech, Language & Hearing Research* 58: 1747–60. [CrossRef]

Bohnacker, Ute, Rima Haddad, and Linnéa Öberg. 2021. Arabic-Swedish-Speaking Children Living in Sweden: Vocabulary Skills in Relation to Age, SES and Language Exposure. *Journal of Home Language Research* 4: 1–18. [CrossRef]

Botting, Nicola, Gina Conti-Ramsden, and Alison Crutchley. 1997. Concordance between Teacher/Therapist Opinion and Formal Language Assessment Scores in Children with Language Impairment. *International Journal of Language & Communication Disorders* 32: 317–27. [CrossRef]

Buac, Milijana, Megan Gross, and Margarita Kaushanskaya. 2014. The Role of Primary Caregiver Vocabulary Knowledge in the Development of Bilingual Children's Vocabulary Skills. *Journal of Speech Language and Hearing Research* 57: 1804–16. [CrossRef] [PubMed]

Calvo, Alejandra, and Ellen Bialystok. 2014. Independent Effects of Bilingualism and Socioeconomic Status on Language Ability and Executive Functioning. *Cognition* 130: 278–88. [CrossRef]

Cartmill, Erica A., Benjamin F. Armstrong, Lila R. Gleitman, Susan Goldin-Meadow, Tamara N. Medina, and John C. Trueswell. 2013. Quality of Early Parent Input Predicts Child Vocabulary 3 Years Later. *Proceedings of the National Academy of Sciences* 110: 11278–83. [CrossRef]

Chiat, Shula, and Penny Roy. 2007. The Preschool Repetition Test: An Evaluation of Performance in Typically Developing and Clinically Referred Children. *Journal of Speech, Language, and Hearing Research* 50: 429–43. [CrossRef]

Chiat, Shula. 2015. Non-Word Repetition. In *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Edited by Sharon Armon-Lotem, Jan de Jong and Natalia Meir. Bristol: Multilingual Matters, pp. 125–47.

Coady, Jeffry A., and Julia L. Evans. 2008. Uses and Interpretations of Non-Word Repetition Tasks in Children with and without Specific Language Impairments (SLI). *International Journal of Language & Communication Disorders* 43: 1–40. [CrossRef]

Cobo-Lewis, Alan, Barbara Pearson, Rebecca Eilers, and Vivian Umbel. 2002a. Effects of Bilingualism and Bilingual Education on Oral and Written English Skills: A Multifactor Study of Standardized Test Outcomes. In *Language and Literacy in Bilingual Children*. Edited by D. Kimbrough Oller and Rebecca Eilers. Clevedon: Multilingual Matters, pp. 64–97.

Cobo-Lewis, Alan, Barbara Pearson, Rebecca Eilers, and Vivian Umbel. 2002b. Effects of Bilingualism and Bilingual Education on Oral and Written Spanish Skills: A Multifactor Study of Standardized Test Outcomes. In *Language and Literacy in Bilingual Children*. Edited by D. Kimbrough Oller and Rebecca Eilers. Clevedon: Multilingual Matters, pp. 98–117.

de Almeida, Laetitia, Sandrine Ferré, Eléonore Morin, Philippe Prévost, Christophe dos Santos, Laurie Tuller, Racha Zebib, and Marie-Anne Barthez. 2017. Identification of Bilingual Children with Specific Language Impairment in France. *Linguistic Approaches to Bilingualism* 7: 331–58. [CrossRef]

dos Santos, Christophe, and Sandrine Ferré. 2018. A Nonword Repetition Task to Assess Bilingual Children's Phonology. *Language Acquisition* 25: 58–71. [CrossRef]

Dollaghan, Chris, and Thomas E. Campbell. 1998. Nonword Repetition and Child Language Impairment. *Journal of Speech, Language, and Hearing Research* 41: 1136–46. [CrossRef]

Dollaghan, Christine A., and Elizabeth A. Horner. 2011. Bilingual Language Assessment: A Meta-Analysis of Diagnostic Accuracy. *Journal of Speech, Language, and Hearing Research* 54: 1077–88. [CrossRef]

Ellis Weismer, Susan, J. Bruce Tomblin, Xuyang Zhang, Paula Buckwalter, Martha Chynoweth, and Maura Jones. 2000. Nonword Repetition Performance in School-Age Children with and without Language Impairment. *Journal of Speech, Language, and Hearing Research* 43: 865–78. [CrossRef] [PubMed]

Friedmann, Naama, and Rama Novogrodsky. 2004. The Acquisition of Relative Clause Comprehension in Hebrew: A Study of SLI and Normal Development. *Journal of Child Language* 31: 661–81. [CrossRef] [PubMed]

Gagarina, Natalia, Sharon Armon-Lotem, Carmit Altman, Zhanna Burstein-Feldman, Annegret Klassert, Nathalie Topaj, Felix Golcher, and Joel Walters. 2014. Age, Input Quantity and Their Effect on Linguistic Performance in the Home and Societal Language among Russian-German and Russian-Hebrew Preschool Children. In *The Challenges of Diaspora Migration: Interdisciplinary Perspectives on Israel and Germany*. Edited by R. K. Silberstein and P. F. Titzmann. Farnham: Ashgate, pp. 63–82.

Ganuza, Natalia, and Christina Hedman. 2019. The Impact of Mother Tongue Instruction on the Development of Biliteracy: Evidence from Somali–Swedish Bilinguals. *Applied Linguistics* 40: 108–31. [CrossRef]

Gathercole, Susan E. 1995. Is Nonword Repetition a Test of Phonological Memory or Long-Term Knowledge? It All Depends on the Nonwords. *Memory & Cognition* 23: 83–94. [CrossRef]

Gathercole, Susan E. 2006. Nonword Repetition and Word Learning: The Nature of the Relationship. *Applied Psycholinguistics* 27: 513–43. [CrossRef]

Gathercole, Susan E., Catherine S. Willis, Alan D. Baddeley, and Hazel Emslie. 1994. The Children's Test of Nonword Repetition: A Test of Phonological Working Memory. *Memory* 2: 103–27. [CrossRef]

Gathercole, Virginia C. Mueller, and Enlli Môn Thomas. 2009. Bilingual First-Language Development: Dominant Language Takeover, Threatened Minority Language Take-Up. *Bilingualism: Language and Cognition* 12: 213–37. [CrossRef]

Gathercole, Virginia C. Mueller, Ivan Kennedy, and Enlli Môn Thomas. 2016. Socioeconomic Level and Bilinguals' Performance on Language and Cognitive Measures. *Bilingualism: Language and Cognition* 19: 1057–78. [CrossRef]

Gibson, Todd A., Connie Summers, Elizabeth D. Peña, Lisa M. Bedore, Ronald B. Gillam, and Thomas M. Bohman. 2015. The Role of Phonological Structure and Experience in Bilingual Children's Nonword Repetition Performance. *Bilingualism: Language and Cognition* 18: 551–60. [CrossRef]

Grimm, Angela, and Petra Schulz. 2014. Specific Language Impairment and Early Second Language Acquisition: The Risk of over- and Underdiagnosis. *Child Indicators Research* 7: 821–41. [CrossRef]

Gutiérrez-Clellen, Vera F., and Gabriela Simon-Cereijido. 2010. Using Nonword Repetition Tasks for the Identification of Language Impairment in Spanish-English-Speaking Children: Does the Language of Assessment Matter? *Learning Disabilities Research & Practice* 25: 48–58. [CrossRef]

Haddad, R. 2017. *Cross-Linguistic Lexical Tasks: Selected Arabic dialects. Adapted from the Lebanese version (CLT-ARA)*, Unpublished material.

Håkansson, Gisela, Eva-Kristina Salameh, and Ulrika Nettelbladt. 2003. Measuring Language Development in Bilingual Children: Swedish-Arabic Children with and without Language Impairment. *Linguistics* 41: 255–88. [CrossRef]

Haman, Ewa, Magdalena Łuniewska, and Barbara Pomiechowska. 2015. Designing Cross-Linguistic Lexical Tasks (CLTs) for Bilingual Preschool Children. In *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Edited by Sharon Armon-Lotem, Jan de Jong and Natalia Meir. Bristol: Multilingual Matters, pp. 196–239.

Haman, Ewa, Magdalena Łuniewska, Pernille Hansen, Hanne Gram Simonsen, Shula Chiat, Jovana Bjekić, Agnė Blažienė, Katarzyna Chyl, Ineta Dabašinskienė, Pascale Engel de Abreu, and et al. 2017. Noun and Verb Knowledge in Monolingual Preschool Children across 17 Languages: Data from Cross-Linguistic Lexical Tasks (LITMUS-CLT). *Clinical Linguistics & Phonetics* 31: 818–43. [CrossRef]

Hamann, Cornelia, and Lina Abed Ibrahim. 2017. Methods for Identifying Specific Language Impairment in Bilingual Populations in Germany. *Frontiers in Communication* 2: 16. [CrossRef]

Hart, Betty, and Todd Risley. 1995. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore: Paul Brookes.

Hoff, Erika, Rosario Rumiche, Andrea Burridge, Krystal M. Ribot, and Stephanie N. Welsh. 2014. Expressive Vocabulary Development in Children from Bilingual and Monolingual Homes: A Longitudinal Study from Two to Four Years. *Early Childhood Research Quarterly* 29: 433–44. [CrossRef]

Holmström, Ketty. 2015. Lexikal Organisation Hos En- och Flerspråkiga Skolbarn med Språkstörning. Doctoral thesis, Lunds Universitet, Logopedics, Phoniatrics and Audiology, Lund, Sweden.

Jones, Gary, Marco Tamburelli, Sarah E. Watson, Fernand Gobet, and Julian M. Pine. 2010. Lexicality and Frequency in Specific Language Impairment: Accuracy and Error Data from Two Nonword Repetition Tests. *Journal of Speech, Language and Hearing Research* 53: 1642–55. [CrossRef]

Kalnak, Nelli, Myriam Peyrard-Janvid, Birgitta Sahlén, and Hans Forssberg. 2012. Family History Interview of a Broad Phenotype in Specific Language Impairment and Matched Controls. *Genes, Brain and Behavior* 11: 921–27. [CrossRef]

Kalnak, Nelli, Myriam Peyrard-Janvid, Hans Forssberg, and Birgitta Sahlén. 2014. Nonword Repetition—A Clinical Marker for Specific Language Impairment in Swedish Associated with Parents' Language-Related Problems. *PLoS ONE* 9: e89544. [CrossRef]

Khoury Aouad Saliby, C., Edith Khouba Hreich, and Camille Messarra. 2017a. *Cross-Linguistic Lexical Tasks: Lebanese Version (CLT-ARA)*, Unpublished material.

Khoury Aouad Saliby, Christel, Christophe dos Santos, Edith Kouba Hreich, and Camille Messarra. 2017b. Assessing Lebanese Bilingual Children: The Use of Cross-Linguistic Lexical Tasks in Lebanese Arabic. *Clinical Linguistics & Phonetics* 31: 874–92. [CrossRef]

Kohnert, Kathryn, Jennifer Windsor, and Dongsun Yim. 2006. Do Language-Based Processing Tasks Separate Children with Language Impairment from Typical Bilinguals? *Learning Disabilities Research & Practice* 21: 19–29. [CrossRef]

Kohnert, Kathryn. 2010. Bilingual Children with Primary Language Impairment: Issues, Evidence and Implications for Clinical Actions. *Journal of Communication Disorders* 43: 456–73. [CrossRef] [PubMed]

Leonard, Laurence B. 2014. *Children with Specific Language Impairment*, 2nd ed. Cambridge: MIT Press.

Leseman, Paul P. M. 2000. Bilingual Vocabulary Development of Turkish Preschoolers in the Netherlands. *Journal of Multilingual and Multicultural Development* 21: 93–112. [CrossRef]

Letts, Carolyn. 2013. What Are the Building Blocks for Language Acquisition? Underlying Principles of Assessment for Language Impairment in the Bilingual Context. In *Solutions for the Assessment of Bilinguals*. Edited by Virginia M. C. Gathercole. Bristol: Multilingual Matters, pp. 36–56.

Lindgren, Josefin, and Ute Bohnacker. 2020. Vocabulary Development in Closely-Related Languages: Age, Word Type and Cognate Facilitation Effects in Bilingual Swedish-German Preschool Children. *Linguistic Approaches to Bilingualism* 10: 587–622. [CrossRef]

Nakagawa, Shinichi, and Holger Schielzeth. 2013. A General and Simple Method for Obtaining R^2 from Generalized Linear Mixed-Effects Models. *Methods in Ecology and Evolution* 4: 133–42. [CrossRef]

Nakagawa, Shinichi, Paul C. D. Johnson, and Holger Schielzeth. 2017. The Coefficient of Determination R^2 and Intra-Class Correlation Coefficient from Generalized Linear Mixed-Effects Models Revisited and Expanded. *Journal of The Royal Society Interface* 14: 1–11. [CrossRef]

National Agency for Education (Skolverket). 2022. Available online: https://www.skolverket.se/skolutveckling/statistik/sok-statistik-om-forskola-skola-och-vuxenutbildning?sok=SokC&verkform=Grundskolan&omrade=Skolor%20och%20elever&lasar=2020/21&run=1 (accessed on 28 March 2022).

Nayeb, Laleh, Thomas Wallby, Monica Westerlund, Eva-Kristina Salameh, and Anna Sarkadi. 2015. Child Healthcare Nurses Believe That Bilingual Children Show Slower Language Development, Simplify Screening Procedures and Delay Referrals. *Acta Paediatrica* 104: 198–205. [CrossRef]

Norbury, Courtenay Frazier, Debbie Gooch, Charlotte Wray, Gillian Baird, Tony Charman, Emily Simonoff, George Vamvakas, and Andrew Pickles. 2016. The Impact of Nonverbal Ability on Prevalence and Clinical Presentation of Language Disorder: Evidence from a Population Study. *Journal of Child Psychology and Psychiatry* 57: 1247–57. [CrossRef]

Öberg, Linnéa. 2020. *Words and Non-Words: Vocabulary and Phonological Working Memory in Arabic-Swedish-Speaking 4–7-Year-Olds with and without a Diagnosis of Developmental Language Disorder*. Studia Linguistica Upsaliensia 27. Uppsala: Acta Universitatis Upsaliensis. Available online: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-421590 (accessed on 15 February 2022).

Ortiz, José A. 2021. Using Nonword Repetition to Identify Language Impairment in Bilingual Children: A Meta-Analysis of Diagnostic Accuracy. *American Journal of Speech-Language Pathology* 30: 2275–95. [CrossRef]

Öztekin, Buket. 2019. *Typical and Atypical Language Development in Turkish-Swedish Bilingual Children Aged 4–7*. Studia Linguistica Upsaliensia 25. Uppsala: Acta Universitatis Upsaliensis.

Paradis, Johanne, and Martha Crago. 2000. Tense and Temporality: A Comparison between Children Learning a Second Language and Children with SLI. *Journal of Speech, Language, and Hearing Research* 43: 834–47. [CrossRef]

Paradis, Johanne, and Theres Grüter. 2014. Introduction to "Input and Experience in Bilingual Development". In *Input and Experience in Bilingual Development*. Edited by Theres Grüter and Johanne Paradis. Trends in Language Acquisition Research 13. Amsterdam: John Benjamins, pp. 1–14.

Paradis, Johanne, Kristyn Emmerzael, and Tamara Sorenson Duncan. 2010. Assessment of English Language Learners: Using Parent Report on First Language Development. *Journal of Communication Disorders* 43: 474–97. [CrossRef]

Paradis, Johanne, Phyllis Schneider, and Tamara Sorenson Duncan. 2013. Discriminating Children with Language Impairment among English-Language Learners from Diverse First-Language Backgrounds. *Journal of Speech, Language and Hearing Research* 56: 971–81. [CrossRef]

Paradis, Johanne, Tamara Sorenson Duncan, Stephanie Thomlinson, and Brian Rusk. 2022. Does the Use of Complex Sentences Differentiate between Bilinguals with and without DLD? Evidence from Conversation and Narrative Tasks. *Frontiers in Education* 6: 804088. [CrossRef]

Parkvall, Mikael. 2016. *Sveriges språk i siffror: Vilka språk talas och av hur många?* Stockholm: Språkrådet & Morfem.

Peña, Elizabeth D., Lisa M. Bedore, and Ellen S. Kester. 2016. Assessment of Language Impairment in Bilingual Children Using Semantic Tasks: Two Languages Classify Better than One. *International Journal of Language & Communication Disorders* 51: 192–202. [CrossRef]

Prevoo, Mariëlle J. L., Maike Malda, Judi Mesman, Rosanneke A. G. Emmen, Nihal Yeniad, Marinus H. Van Ijzendoorn, and Mariëlle Linting. 2014. Predicting Ethnic Minority Children's Vocabulary from Socioeconomic Status, Maternal Language and Home Reading Input: Different Pathways for Host and Ethnic Language. *Journal of Child Language* 41: 963–84. [CrossRef] [PubMed]

Purse, Katie, and Hilary Gardner. 2013. Does Formal Assessment of Comprehension by SLT Agree with Teachers' Perceptions of Functional Comprehension Skills in the Classroom? *Child Language Teaching and Therapy* 29: 343–57. [CrossRef]

R Core Team. 2021. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Vienna: R Core Team. Available online: https://www.R-project.org/ (accessed on 15 February 2022).

Radeborg, Karl, Eleonora Barthelom, Monika Sjöberg, and Birgitta Sahlén. 2006. A Swedish Non-Word Repetition Test for Preschool Children. *Scandinavian Journal of Psychology* 47: 187–92. [CrossRef]

Restrepo, María Adelaida. 1998. Identifiers of Predominantly Spanish-Speaking Children with Language Impairment. *Journal of Speech, Language, and Hearing Research* 41: 1398–411. [CrossRef]

Reuterskiöld, Christina, Anna Eva Hallin, Vishnu K. K. Nair, and Kristina Hansson. 2021. Morphosyntactic Challenges for Swedish-Speaking Children with Developmental Language Disorder in Comparison with L1 and L2 Peers. *Applied Linguistics* 42: 720–39. [CrossRef]

Rice, Mabel L., and Lesa Hoffman. 2015. Predicting Vocabulary Growth in Children with and without Specific Language Impairment: A Longitudinal Study from 2;6 to 21 Years of Age. *Journal of Speech, Language and Hearing Research* 58: 345–59. [CrossRef]

Ringblom, Natasha, Gisela Håkansson, and Josefin Lindgren. 2014. *Cross-Linguistic Lexical Tasks: Swedish version (CLT-SWE)*, Unpublished material.

Rowe, Meredith L. 2012. A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development* 83: 1762–74. [CrossRef]

Sahlén, Birgitta, Christina Reuterskiold-Wagner, Ulrika Nettelbladt, and Karl Radeborg. 1999. Non-Word Repetition in Children with Language Impairment—Pitfalls and Possibilities. *International Journal of Language & Communication Disorders* 34: 337–52. [CrossRef]

Salameh, Eva-Kristina, Gisela Håkansson, and Ulrika Nettelbladt. 2004. Developmental Perspectives on Bilingual Swedish-Arabic Children with and without Language Impairment: A Longitudinal Study. *International Journal of Language & Communication Disorders* 39: 65–90. [CrossRef]

Salameh, Eva-Kristina, Ulrika Nettelbladt, and B. Gullberg. 2002. Risk Factors for Language Impairment in Swedish Bilingual and Monolingual Children Relative to Severity. *Acta Paediatrica* 91: 1379–84. [CrossRef] [PubMed]

Schwob, Salomé, Laurane Eddé, Laure Jacquin, Mégane Leboulanger, Margot Picard, Patricia Ramos Oliveira, and Katrin Skoruppa. 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: 3578–93. [CrossRef] [PubMed]

Smolander, Sini, Marja Laasonen, Eva Arkkila, Pekka Lahti-Nuuttila, and Sari Kunnari. 2021. L2 Vocabulary Acquisition of Early Sequentially Bilingual Children with TD and DLD Affected Differently by Exposure and Age of Onset. *International Journal of Language & Communication Disorders* 56: 72–89. [CrossRef]

Sorenson Duncan, Tamara, and Johanne Paradis. 2016. English Language Learners' Nonword Repetition Performance: The Influence of Age, L2 Vocabulary Size, Length of L2 Exposure, and L1 Phonology. *Journal of Speech, Language and Hearing Research* 59: 39–48. [CrossRef] [PubMed]

SOU. 2016. Samordning, Ansvar och Kommunikation: Vägen till Ökad Kvalitet i Utbildningen För Elever med Vissa Funktionsnedsättningar: Slutbetänkande. Available online: https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2016/06/sou-201646/ (accessed on 4 June 2020).

Spaulding, Tammie J., Sabrina Hosmer, and Calli Schechtman. 2013. Investigating the Interchangeability and Diagnostic Utility of the PPVT-III and PPVT-IV for Children with and without SLI. *International Journal of Speech-Language Pathology* 15: 453–62. [CrossRef] [PubMed]

Thordardottir, Elin, and Myrto Brandeker. 2013. The Effect of Bilingual Exposure versus Language Impairment on Nonword Repetition and Sentence Imitation Scores. *Journal of Communication Disorders* 46: 1–16. [CrossRef]

Thordardottir, Elin. 2011. The Relationship between Bilingual Exposure and Vocabulary Development. *International Journal of Bilingualism* 15: 426–45. [CrossRef]

Thordardottir, Elin. 2015. Proposed Diagnostic Procedures for Use in Bilingual and Cross-Linguistic Contexts. In *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Edited by Sharon Armon-Lotem, Jan de Jong and Natalia Meir. Bristol: Multilingual Matters, pp. 331–58.

Thordardottir, Elin. 2019. Amount Trumps Timing in Bilingual Vocabulary Acquisition: Effects of Input in Simultaneous and Sequential School-Age Bilinguals. *International Journal of Bilingualism* 23: 236–55. [CrossRef]

Tomblin, J. Bruce, Nancy L. Records, Paula Buckwalter, Xuyang Zhang, Elaine Smith, and Marlea O'Brien. 1997. Prevalence of Specific Language Impairment in Kindergarten Children. *Journal of Speech, Language, and Hearing Research* 40: 1245–60. [CrossRef] [PubMed]

Trauner, Doris, Beverly Wulfeck, Paula Tallal, and John Hesselink. 2000. Neurological and MRI Profiles of Children with Developmental Language Impairment. *Developmental Medicine & Child Neurology* 42: 470–75. [CrossRef]

Tuller, Laurice, Cornelia Hamann, Solveig Chilla, Sandrine Ferré, Eléonore Morin, Philippe Prevost, Christophe dos Santos, Lina Abed Ibrahim, and Racha Zebib. 2018. Identifying Language Impairment in Bilingual Children in France and in Germany. *International Journal of Language & Communication Disorders* 53: 888–904. [CrossRef]

Tuller, Laurie. 2015. Clinical Use of Parental Questionnaires in Multilingual Contexts. In *Methods for Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Edited by Sharon Armon-Lotem, Jan de Jong and Natalia Meir. Bristol: Multilingual Matters, pp. 301–30.

UNESCO Institute for Statistics. 2012. *International Standard Classification of Education: ISCED 2011*. Montreal: UNESCO Institute for Statistics. Available online: http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf (accessed on 15 February 2022).

Unsworth, Sharon. 2016. Early Child L2 Acquisition: Age or Input Effects? Neither, or Both? *Journal of Child Language* 43: 608–34. [CrossRef] [PubMed]

World Health Organization. 1992. *ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.